|NST|TUTO
UN|VERS|TÁR|O
DE L|SBOA

Análise da propagação e impacto da informação nas redes sociais

Miguel Dordio Lobo da Conceição Oliveira

Mestrado em Engenharia Informática

Orientadores:
Doutor Nuno Manuel de Carvalho Ferreira Guimarães,
Professor Catedrático,
ISCTE-IUL

Doutor António Jorge Filipe da Fonseca, Professor Convidado,
ISCTE-IUL

setembro, 2022

TECNOLOGIAS
E ARQUITETURA

Departamento de Ciências e Tecnologia da Informação

Análise da propagação e impacto da informação nas redes sociais

Miguel Dordio Lobo da Conceição Oliveira

Mestrado em Engenharia Informática

Orientadores:
Doutor Nuno Manuel de Carvalho Ferreira Guimarães,
Professor Catedrático,
ISCTE-IUL

Doutor António Jorge Filipe da Fonseca, Professor Convidado,
ISCTE-IUL

setembro, 2022

Agradecimentos

A conclusão desta dissertação vem assim representar o fim de uma etapa de grande superação pessoal, onde tive a oportunidade de testar os meus limites enquanto superava desafios profissionais ao mesmo tempo que académicos, alcançada com a continua determinação, mas também o apoio incondicional das pessoas que me acompanharam neste processo.

Quero assim deixar os meus sinceros agradecimentos aos meus pais, Paula e Luís, que sempre me apoiaram e incentivaram ao longo do meu percurso académico, providenciando continuamente qualquer ajuda necessária.

À minha namorada, Sofia, me tem vindo a acompanhar durante o meu ensino superior, sempre com paciência e dedicação, capaz de me fazer ver que tudo é possível com o devido esforço.

À minha irmã, Maria, que sempre se disponibilizou para prestar apoio e conselhos valiosos nos momentos de revisão do trabalho feito. Mostrou-se sempre disponível, nas horas de maior dificuldade, a fim de me ajudar quando eu precisava de pôr as ideias em ordem.

Aos meus avós, que me têm vindo a acompanhar e apoiar durante todos os capítulos da vida e que me vêm, agora, a chegar ao fim de mais uma etapa.

Por fim, quero agradecer aos meus orientadores, professor Nuno e professor António, por me direcionarem durante este processo, mostrando-se sempre pró-ativos e disponíveis para me encaminharem rumo ao destino certo, continuamente dando conselhos assertivos para levar a dissertação a bom porto.

Resumo

Cada vez mais o uso de redes sociais é uma constante na sociedade moderna, o que leva a que todos os dias, uma grande quantidade de informação seja criada pelos seus utilizadores. Deste modo, e com os mais variados tópicos a serem divulgados, noticiados e debatidos nestas plataformas, este estudo pretende caracterizar um processo contínuo de recolha, processamento, análise e previsão de vetores de popularidade dentro das redes sociais. Usaremos o Twitter como base de investigação.

Esta investigação foca-se em desenhar uma metodologia capaz de recolher uma amostra do conteúdo circulante no Twitter durante um dado período de tempo, o mais fielmente possível, recorrendo ao uso das ferramentas disponíveis, nomeadamente a API do Twitter. Para além disso, é ainda apresentado um processo de recolha de utilizadores (e suas informações) que partilharam os tweets recolhidos. Este processo surge como uma alternativa mais versátil à atualmente fornecida oficialmente pela API do Twitter que limita consideravelmente os dados possíveis de recolher nesta vertente.

É também apresentada uma proposta de processamento dos dados recolhidos de forma a extrair, a análises à base de gráficos e tabelas de forma automática, de forma a mais facilmente ilustrar o comportamento de elementos de informação, finalizando no desenvolvimento de um modelo capaz de prever se um tweet será ou não popular.

Palavras-Chave: Popularidade, Disseminação, Automatização, Machine Learning, Twitter

Abstract

Increasingly the use of social networks is a constant in modern society, which leads that every day, a large amount of information is created by its users. Thus, and with the most varied topics being disseminated, reported, and debated on these platforms, this study aims to characterize a continuous process of collection, processing, analysis, and prediction of popularity vectors within social networks. We will use Twitter as the research base.

This research focuses on designing a methodology capable of collecting a sample of the content circulating on Twitter during a given period, as faithfully as possible, through the use of available tools, namely the Twitter API. In addition, a process for collecting users (and their information) who shared the collected tweets is also presented. This process appears as a more versatile alternative to the one currently officially provided by the Twitter API, which considerably limits the data that can be collected in this aspect.

It is also presented a proposal for processing the collected data in order to extract, to graph and table-based analysis in an automatic way, in order to more easily illustrate the behavior of information elements, ending in the development of a model capable of predicting whether or not a tweet will be popular.

Keywords: Popularity, Dissemination, Automation, Machine Learning, Twitter.

Índice Geral

Índice de Tabelas

Índice de Equações

Índice de Figuras

Glossário de Abreviaturas e Siglas

API - Application programming interface

DSR - Design Science Research

LDA – Algoritmo Latent Dirichlet Allocation

LED - Algoritmo Loop Edges Delete

NOVER - Algoritmo Neighborhood overlap

VADER - Valence Aware Dictionary and Sentiment Reasoner

BCP – Sistema de códigos para idiomas

Capítulo 1 – Introdução

1.1. Motivação e enquadramento

A crescente tendência para a digitalização da realidade e para a comunicação por meios digitais, introduz diariamente grandes quantidades de dados e informação nas plataformas digitais. Entre elas, existem as redes sociais, plataformas que na sua ideologia mais primária, pretendem conectar pessoas de forma fácil, rápida e gratuita.

Com o aumento de popularidade destas redes sociais, este intuito inicial tem sido expandido gradualmente, levando o propósito inicial das mesmas a divergir da simples conexão entre pessoas. Agora, temas do passado, atualidade e até futuro são amplamente discutidos e cada vez mais estas plataformas apresentam um vasto conteúdo opinativo, factual, falso e enganador, tudo misturado num só local.

Toda esta informação disponível suscita cada vez mais estudos que pretendem compreender e até modular os comportamentos diariamente demonstrados pelas mais diversas pessoas, dos mais variados espetros sociais e geopolíticos.

Assim, surge igualmente a necessidade de tentar perceber o comportamento da informação circulante nestas redes, ou seja, analisar o fluxo, meio de propagação, dispersão e o resultado final desse comportamento, isto é, o impacto que uma informação circulante nas redes sociais pode ter no mundo "real".

Para isto, são normalmente analisadas as principais redes sociais no mundo ocidental (exclui-se o espaço asiático, com plataformas e redes sociais próprias), como o Twitter, rede social que se caracteriza pelo seu formato de mensagens curtas intituladas de "tweets", agrupamento de conteúdo por "hashtags" e onde uma vasta quantidade de líderes políticos interage com os seus cidadãos.

A disponibilização, por exemplo, imediata do conteúdo por parte das redes sociais, torna-o facilmente partilhado em tempo real pelos utilizadores. Isto tem como consequência que muito do conteúdo ali presente seja de caracter opinativo, que normalmente contém vários erros ortográficos, palavras abreviadas, siglas e acrónimos. Deste modo, ferramentas de processamento textual, sanitização e classificação têm que ser usadas e customizadas para poder obter resultados neste contexto.

Para além disso, diferentes abordagens podem ser usadas para analisar a propagação de informação nestas redes, desde algoritmos de análise de sentimento à identificação de

comunidades por onde a informação circula ou que se formam em torno de certos tópicos. Para isto, metodologias e ferramentas já bem estabelecidas na área podem ser aplicadas.

1.2. Questões de investigação

No âmbito do tema em estudo, as questões de investigação que motivam a análise elaborada encontram-se apresentadas a seguir:

- Quais são as técnicas mais eficazes para de modelar o comportamento da propagação de informação nas redes sociais?
- Quais os padrões e características comuns que uma informação deve/pode ter para maximizar a sua disseminação nas redes sociais?
- Pode a amplitude e o alcance de uma dada informação ser previsto à priori?
- O tópico e ou sentimento de uma publicação numa rede social tem impacto na sua disseminação?
- Existe um perfil de utilizador padrão que consegue que o seu conteúdo seja mais divulgado?
- Quais são as características comuns a utilizadores que partilham conteúdo nas redes sociais?
- Quais os períodos do dia/semana em que a informação tem maior alcance?

1.3. Objetivos

O objetivo desta dissertação é a conceção e desenvolvimento de ferramentas computacionais para extrair redes de elementos de informação relacionados e analisar a sua estrutura. Como problema base, são consideradas as notícias publicadas numa rede social "Twitter" e todas as ações consequentes – "retweets", comentários, "likes", etc.

Neste trabalho, pretende-se construir uma cadeia de ferramentas, começando na conceção de um sistema capaz de extrair automaticamente uma amostra, o mais representativa possível, de uma rede social (Twitter) para um dado período de tempo. Posteriormente será aplicado um tratamento aos dados armazenados, de modo, a garantir que se encontram num formato acessível de processamento. De seguida, o sistema será ainda capaz, de automaticamente, criar representações (recorrendo a bibliotecas auxiliares de gráficos e visualização de dados) apropriados a, de maneira sucinta, ilustrarem o estado atual do comportamento nas redes sociais, de acordo com a informação recolhida, bem como a caracterização dos utilizadores que partilham e criam conteúdo nas redes sociais. Por fim, propõe-se a criação de um modelo que faz uma seleção automática dos melhores dados recolhidos e é capaz de prever com uma certa percentagem de precisão se uma dada publicação numa rede social será, ou não, popular.

Em resumo a dissertação pretende modelar o comportamento dinâmico e fluído da informação circulante nas redes sociais, recorrendo à conceção e implementação de software capaz de lidar com dados na sua forma original e deles derivar padrões de comportamento, para que seja possível melhor compreender o alcance e o impacto.

1.4. Processo de investigação

O processo de investigação, no quadro dos objetivos delineados, baseia-se no modelo de Design Science Research (DSR). A metodologia seguida por este modelo tem como sua base a construção e avaliação de relações genéricas meio/fim, sendo na sua essência um processo de resolução de problemas (Peffers et al., 2007).

Este processo é assim caracterizado por seis etapas que em conjunto irão conduzir à finalização do protótipo a ser desenvolvido e sua avaliação final, podendo ser esquematizado da seguinte forma:

1ª Fase: **Identificação do problema e o seu motivo**, onde se procura desenvolver um artefacto que define o problema a ser investigado bem como o valor da sua solução,

2ª Fase: **Definição de objetivo**, onde é feita a derivação de propósitos e fins a alcançar de uma solução, sendo que estes podem ser quantitativos ou qualitativos,

3ª Fase: **Desenho e desenvolvimento**, realização da arquitetura e conceção do sistema computacional que irá modelar os comportamentos que se tentam prever, seguindo uma abordagem de desenvolvimento incremental, deixando o sistema cada vez mais robusto,

4ª Fase: **Demonstração**, onde é efetuada a quantificação do protótipo desenvolvido, de modo a aferir a sua capacidade de dar resposta ao problema que se pretende resolver, que através de testes e simulações deve ser capaz de apresentar resultados finais,

5ª Fase: **Avaliação**, do desenvolvimento realizado e a sua demonstração, de forma a quantificar o quão eficaz a investigação em comparação com os objetivos inicialmente propostos.

6ª Fase: **Comunicação**, que culmina (neste caso) na escrita e apresentação da investigação desenvolvida sob a forma de uma dissertação que expõe os processos incrementais que levaram à identificação das respostas ao tema.

Este método cíclico de desenvolvimento, pressupõe a melhoria incremental do processo em causa onde as repetições das fases acima enumeradas serão revistas sempre que necessário, culminando no atingir dos resultados definidos no início da investigação e a sua exposição.

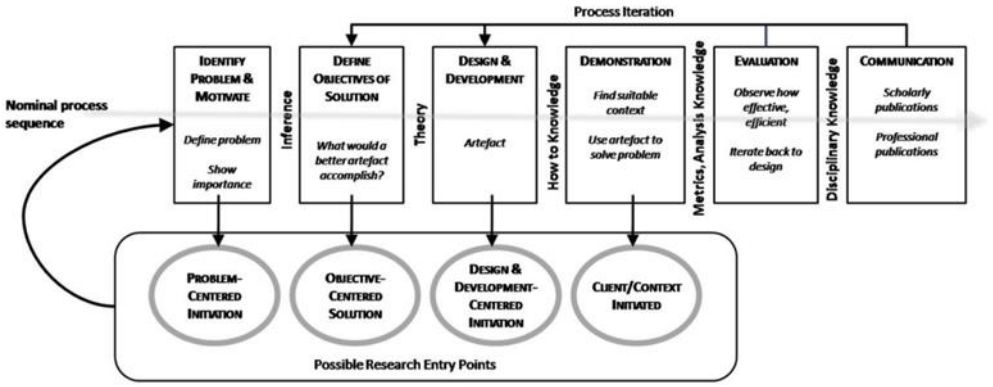

Figura 1: Metodologia Design Science Research (Peffers et al., 2007)

1.5. Estrutura e organização da dissertação

O presente trabalho encontra-se organizado em cinco capítulos que procuram espelhar as diferentes fases executadas até à sua conclusão. O primeiro capítulo introduz o tema da investigação e seus objetivos, bem como uma descrição da metodologia e estrutura utilizadas durante o processo. O segundo capítulo apresenta a revisão de literatura, que reflete um enquadramento teórico do tema. O terceiro capítulo é dedicado ao procedimento usado no processo de recolha e tratamento dos dados, bem como à definição do sistema desenvolvido para suporte destas funções. O capítulo quatro apresenta as análises dos resultados obtidos. Por fim, o capítulo cinco relata as conclusões desta investigação, bem como recomendações, limitações e trabalho futuro.

Capítulo 2 – Revisão da Literatura

2.1. O que é uma rede social

O modo de vida de um indivíduo é em grande parte influenciado e derivado das conexões que este estabelece em sociedade. Deste modo, uma rede social começou por ser um local onde indivíduos podiam estender as suas interações mesmo quando não estão juntos, potenciando meios de interação diferentes das tradicionais chamadas de voz ou mensagens curtas (sms).

As redes sociais responderam a uma crescente vontade de pessoas, dos mais diversos espetros sociais, em partilhar as suas vidas pessoais para com o mundo, levando à constante procura por validação aliada à necessidade de exibir aspetos das suas vidas pessoais.

Contudo, olhando para lá de todas as vidas pessoais, realça-se a facilidade com que as redes sociais conseguem gerar debates e trocas de ideias em temas centrais, nas mais diferentes comunidades, levando à criação de um espaço onde os indivíduos se sentem livres de expressar o que pensam sobre as mais variadas matérias. Isto atribui às redes sociais um papel importante na sociedade moderna, onde todos os temas emergentes, ou até mesmo temas perdidos no passado, rapidamente se tornam virais e alvo de debate durante dias.

Com o aparecimento das redes sociais, o seu rápido crescimento e a sua propagação nas últimas duas décadas, inúmeros estudos têm sido realizados, com o intuito de compreender a natureza e dinâmica do seu funcionamento, bem como os comportamentos evidenciados pelos seus utilizadores, incidindo particularmente no modo como ideias ou notícias surgem e são amplificadas por estes meios.

2.2. Importância e crescimento das redes sociais

A crescente quantidade de dados gerada por redes sociais tem revelado uma grande importância no contributo para alcançar os objetivos de sustentabilidade e que melhoram a sociedade. Este fenómeno é demonstrado pelo investimento por parte de organizações, empresas e governos em políticas mais sustentáveis, tais como direitos das mulheres, sustentabilidade ambiental e saúde, fruto de campanhas virais nas redes sociais (Can & Alatas, 2017).

As redes sociais vêm também providenciar um meio rápido, e conveniente de acesso a notícias. Nunca antes foi tão fácil conseguir obter informação e ficar a par do que está a acontecer, seja a uma escala pequena, como acontecimentos locais, às grandes notícias mundiais do momento. Assim, cada vez mais se torna palpável o impacto e influência das redes sociais para além do mundo digital.

À medida que vão ganhando popularidade, as redes sociais, têm inadvertidamente, impactado o "mundo real", através das ideias que nelas se materializam ou até funcionando como um catalisador, capaz de impulsionar e modificar o resultado de eventos como a primavera árabe ou eleições dos mais variados governos pelo mundo (Brady et al., 2017).

2.3. Propagação de ideias

Apesar da grande variedade de informação em circulação nas redes sociais, as opiniões sensacionalistas conseguem, eficazmente, alcançar o pretendido, captar a atenção de quem as ouve, lê ou vê. Com base nisto, mensagens com as quais se pretende atingir a maior audiência possível, tendem a ser o mais sensacionalistas possível, por vezes, usando títulos dúbios e de grande impacto para rapidamente captarem a atenção. Isto é evidente quando mensagens contendo palavras de teor moral/emocional têm 20% mais probabilidade de alcançar um maior público (Brady et al., 2017).

Tendo este facto em conta, é natural que figuras públicas recorram a este tipo de linguagem quando se expressam nestas plataformas digitais de maneira a conseguirem apelar à sua audiência, pois conhecem o poder que o apelo ao lado emocional do ser humano consegue ter. Assim, é importante investigar que tipo de emoções está ligado a certos resultados para melhor compreender a propagação de ideias em redes sociais.

2.4. Desinformação nas redes sociais

A dispersão de informação falaciosa, falsa, tendenciosa ou parcialmente incorreta tem vindo a aumentar significativamente nos últimos anos, assumindo as mais diversas formas como conspirações, notícias fabricadas, rumores sem fundamentação, etc.

Este tipo de informação pode ser classificado como "desinformação" e caracteriza-se como informação que não é verdadeira, mas para a qual é incerta a intenção que levou à sua divulgação. Por outro lado, existem também os "rumores" que consistem em informações cuja legitimidade não foi ainda confirmada.

Existe ainda a denominação de "fake news" ou até "trolling" que diverge da desinformação, por ter intenções maliciosas na sua origem, procurando criar controvérsias e reações emocionais ou até ofender (Binns, 2012).

Analisando a origem, e consequente, propagação desta desinformação, demonstra-se que apesar de esta ter início e formatação em sites "clickbait" ou até sites partidários, a sua divulgação é feita por meio de "influencers", ou seja, utilizadores com grandes redes de seguidores que, ao partilharem estes conteúdos, os direcionam rapidamente a uma grande audiência. Este padrão repete-se apresentando sempre a mesma ideia com um formato ligeiramente alterado até que o assunto deixe de ser mediático (Shin et al., 2018).

A exposição a este tipo de conteúdo pode, como tem vindo a ser demonstrado, alterar as crenças de um indivíduo e, por conseguinte, a sua maneira de agir e as ações que tem para consigo ou pessoas em seu redor (Shin et al., 2018), podendo até ter consequências graves, quando se trata de temas políticos ou até de saúde (Waszak et al., 2018).

Ao procurar identificar o tipo de utilizador mais propício de divulgação de informação falsa, observa-se que são os utilizadores com menos seguidores que têm maior probabilidade de o ser, e que para além disso são utilizadores que por norma têm pouca atividade e presença nas redes sociais (Vosoughi et al., 2018).

2.5. Fragmentação de ideologias nas redes sociais

Para além da preocupação com a desinformação e propaganda presente nas redes sociais, existe ainda a preocupação da convergência de pessoas em torno de uma ideologia, normalmente de relevo político, mas sem exclusividade. Isto pode levar à formação de comunidades "fechadas", que proporcionam um ambiente "seguro" e distante para os seus membros de opiniões e perspetivas contrárias, este fenómeno tem o nome de "bolhas" / "echo chambers" (Knobloch-Westerwick & Meng, 2009).

Sendo as redes sociais um ambiente propício para este tipo de situações, é importante notar que a redução de comunicações entre formas de pensar diferentes leva a que fragmentação dentro das plataformas acentue. Apesar disso, é possível que as ideologias tenham relevância no resultado desta acentuação, bem como o tamanho do próprio grupo que se forma em torno de um dado tópico.

Existem várias abordagens possíveis para a análise deste fenómeno como, por exemplo, o acompanhamento do desenrolar de um "hashtag" no Twitter aliada à criação de uma métrica para medir o nível de fragmentação. A adição de análise de sentimento ao conteúdo das mensagens trocadas, poderia levar a uma maior precisão nos resultados, bem como a incorporação do contexto temporal de cada situação (Bright, 2018).

2.6. Técnicas de análise

2.6.1. Recolha de dados

Para fomentar e suster os estudos feitos na área, os investigadores usam APIs disponibilizadas diretamente pelas redes sociais onde o estudo será conduzido a fim de recolher os dados necessários. A API mais popular é a do Twitter que disponibiliza vários tipos de acesso aos seus dados, oferecendo uma versão mais relaxada nos seus limites a investigadores académicos.

Contudo a existência de planos grátis e pagos já relevou em estudos anteriores que os dados obtidos são diferentes entre os planos, produzindo resultados diferentes. (Antonakaki et al., 2021).

A API do twitter fornece os dados em formato JSON e devido à sua natureza mais complexa é observada uma tendência no uso de bases de dados que suportam este formato à priori como bases de dados NoSQL em prol das convencionais que usam modelos relacionais (Antonakaki et al., 2021).

2.6.2. Análise linguística

Quando se pretende elaborar uma análise linguística, isto é, uma análise onde o conteúdo e contexto da mensagem transmitida num texto é relevante para a investigação, é preciso previamente garantir que o texto é na melhor das capacidades “limpo”, ou seja, que sejam removidos elementos como “hashtags”, URLs, e mensagens demasiado curtas, a fim de garantir maior eficácia para as ferramentas de análise.

De modo a fazer este tipo de análise, existem ferramentas como o VADER (Valence Aware Dictionary and Sentiment Reasoner) que se baseia em léxicos de palavras relacionadas a sentimentos, permitindo a classificação automática de cada palavra no léxico como positiva, neutra ou negativa e conseguindo ainda inferir o nível do sentimento (Hung et al., 2020). Para além deste, o modelo de LDA (Latent Dirichlet Allocation) é capaz de lidar com textos curtos ou longos, onde agrupa as palavras comuns em múltiplos tópicos, que o torna apropriado para extrair as estruturas semânticas presentes num dado texto, permitindo que se consiga classificar a similaridade de temas (A. Yang et al., 2021).

O processo de Lemmatization, reduz formas flexionais das palavras a uma raiz comum ou a um único termo, e consegue identificar, dentro de um conjunto de palavras, qual o termo comum, normalmente na sua forma de dicionário, funcionado assim como um processo de normalização de texto (Liu et al., 2012).

Recorrendo a processos de análise de texto com mecanismos de inteligência artificial, aprendizagem automática e mineração de texto, é possível utilizar algoritmos capazes de análise de sentimento. Esta análise é útil para melhor compreender a perceção do público geral em torno de um determinado tema (A. Yang et al., 2021).

2.6.3. Identificação de comunidades

Para além destes métodos de classificação de conteúdo é necessário, normalmente, um meio de agrupamento e identificação de "comunidades", isto é, identificação do grupo de indivíduos envolvidos com um dado tópico. A deteção de comunidades possibilita a compreensão da estrutura topológica e previsão do comportamento de redes complexas.

Existem diversos algoritmos de identificação de comunidades. Tal como o algoritmo LED, que é bastante eficiente a encontrar comunidades, conseguindo manter uma complexidade temporal linear, mesmo em grandes redes. Ainda no mesmo formato existe o algoritmo NOVER, que alcança o mesmo objetivo, mas tendo a sobreposição de vizinhos como heurística para a importância das arestas/ligações do grafo. Ambos os algoritmos usam a força dos seus nós para remover arestas (X. Chen & Li, 2019).

Outros algoritmos usam a relação entre nós para remoção de arestas, como o algoritmo de agrupamento hierárquico, que procura a similaridade entre todos os pares de nós, levando à fusão num só, ou ainda o algoritmo de Louvain, cujo objetivo é maximizar a modularidade de uma rede, com a desvantagem de poder falhar em redes de pequena dimensão. O algoritmo de LPA, procura similaridades locais, de modo a agrupar membros da comunidade. Quanto a algoritmos com foco no método de remoção de arestas, temos o algoritmo EDCD, que apresenta maior desempenho a lidar com redes e produz comunidades de maior qualidade (X. Chen & Li, 2019).

Um outro método clássico de identificação de comunidades é o algoritmo aglomerativo, que tende a originar uma espécie de árvore hierárquica para representar os grupos dos componentes (Li et al., 2019). Com isto, pode-se ainda acrescentar uma

análise à estrutura de ligação dentro do grafo, recorrendo ao algoritmo PageRank, com o qual se consegue calcular a possível importância de um dado nó.

A análise de estruturas de redes sociais é um processo que requer diferentes soluções nas várias etapas de desenvolvimento, mas que pode ser unificado num só fluxo de trabalho para tornar o processo automático.

Este fluxo de trabalho é caracterizado pelo armazenamento de dados diretamente em bases de dados adequadas para grafos a fim de melhorar a velocidade e facilidade de filtrar para agregação. Na fase de análise deve ser tida em conta a topologia do grafo e o recurso a algoritmos próprios. No fim, o grafo pode ser visualizado com ferramentas apropriadas (Kolomeets et al., 2019).

2.6.4. Visual

A abordagem de classificação visual, procura avaliar a classificação de claridade, coerência, agrupamento e diversidade. Já de uma perspetiva estatística, propriedades como o rácio de imagem, multi-imagens, são pontos interessantes a ter em conta (Shu et al., 2017). Assim, notícias tendem a conter imagens sensacionais de modo a obter uma reação emocional.

Deste modo pode-se depois analisar as notícias numa perspetiva de utilizador, de publicações ou de redes, isto é, procurando por características de utilizadores que interagem em redes sociais, podendo ser subdivididas a um nível individual ou de grupo, por características de determinadas publicações, como o tópico, credibilidade e posicionamento; ou procurando por redes formadas por utilizadores que postam conteúdos similares nas redes sociais.

Para a análise de dados, diferentes metodologias podem ser seguidas. Assim, para analisar tópicos salientes, boas abordagens seriam, nuvens de palavras, para partilha de recursos e links, diagramas de Euler seriam apropriados para respostas de cidadãos, diagramas de Hasse ou a roda das emoções de Plutchik’s, uma boa aposta para a perceção de sentimentos e, para análise de emoções, gráficos “bolha” seriam uma escolha viável (Hubert et al., 2020).

Gráficos bolha são bons para identificar a variação de intensidade e polaridade ao longo do tempo. Diagramas de Euler e Sankey são boas opções para identificar estratégias

de comunicações, enquanto diagramas de Hasse estendido são bons para detetar a maneira como as pessoas reagem (Hubert et al., 2020).

2.6.5. "Bots" e "Spam"

Um fator de grande relevância nas redes sociais é a existência de "bots" (programas autónomos capazes de interagir com outros sistemas ou utilizadores) e "spam" (mensagens não solicitadas ou irrelevantes enviadas pela internet) nas redes sociais que tentam deturpar o uso normal das mesmas pelos seus utilizadores. Assim surge a necessidade de sistemas de identificação e classificação de spam e bots.

O processo de identificação de "spam" segue um fluxo de normal de análise de dados, usando vários conceitos e algoritmos de "machine learning". Deste modo existem três tipos de classificação possível: deteção de spam por tweets, por utilizadores e por campanhas (Chu et al., 2012).

Por outro lado, um método também amplamente usado é o recurso a inspetores humanos. Este método consegue um resultado de classificação superior, contudo tem a desvantagem de exigir grande esforço humano (Benevenuto et al., 2010).

Complementado a identificação de spam, a presença de contas falsas com os mais diversos propósitos levava a necessidade de ferramentas de deteção de bots. Esta é uma tarefa difícil, sendo uma das ferramentas mais conhecida, o botornot (Davis et al., 2016) que mais tarde se desenvolveu no Botometer (K. C. Yang et al., 2019).

2.7. Conclusão

As redes socias têm diversas estratégias de difusão de informação, cujo objetivo é aumentar o número de interações, visualizações e ultimamente maximizar o tempo do utilizador na rede. Assim estratégias como procurar apresentar aos utilizadores informações e contactos que não lhes são muito próximos ou até nem existem, potenciando a expansão e interação com novos temas e pessoas, ou até, exibindo novos conteúdos partilhados por outros utilizadores da rede de contactos/interesse.

As informações que mais visibilidade têm são as que mais provavelmente se tornaram virais. Contudo, é curioso notar que por norma não são as grandes contas de influencers que criam aquilo que serão os temas da atualidade. Ao contrário do que se pode pensar, a informação viral não se espalha como uma doença patogénica, em que uma grande rede de contacto proporciona melhores condições de suceder, pois pessoas nestas posições, tendem a ficar sobrecarregadas de informação, levando a que a atenção diminua quando há tanto para ver (Hodas & Lerman, 2013).

Muitos estudos que são feitos, apesar de incorporarem na sua análise diferentes redes sociais, tendem a combinar as métricas numa variável, descartando a heterogeneidade de cada plataforma. Para além disso, a maioria dos estudos baseia-se em pesquisas transversais, que não conseguem dar resposta ao problema da casualidade ou prendem-se em análises que incidem maioritariamente numa população jovem, que acaba por não representar realisticamente a população geral (Halpern et al., 2017).

Assim é importante continuar a analisar o comportamento e dinamismos característicos de redes sociais de modo a compreender de melhor forma como a informação se propaga e circula dentro destas redes para assim perceber de que modo isso pode afetar e se traduz para fora delas. Uma vez que nunca antes, o que acontece no mundo digital teve tanto impacto no “real” como nos dias de hoje.

Capítulo 3 – Construção de uma amostra de dados sobre o Twitter

3.1. Twitter

Neste capítulo é feita uma descrição do funcionamento da rede social Twitter, usada como base da investigação e de onde foram coletados os dados analisados. Seguidamente apresentamos a explicação do funcionamento da API do Twitter, ferramenta usada na recolha dos dados, e uma exposição dos dados recolhidos, bem como do processo de recolha, culminando na caracterização do tratamento feito posteriormente.

A rede social Twitter pode ser usada das mais variadas formas e com os mais variados motivos. Esta versatilidade permite a captação de um vasto leque de utilizadores da rede social, das mais variadas origens, desde a sua utilização mais casual até ao outro extremo do espetro, como um veículo de transmissão de informação oficial, como se vê, pela elevada quantidade de publicações por oficiais de governos de todo o mundo.

O Twitter destaca-se das outras redes sociais pela forma característica como permite aos utilizadores partilharem informação, através de texto escrito. Contudo este formato impõe uma restrição de caracteres máximos por cada publicação, que foi expandida de um limite de 140 caracteres para 280 (desde novembro de 2017). Estas condições conduzem os utilizadores para um formato de escrita mais simples, sem prestar muita atenção à formatação do texto, mas focando-se mais no conteúdo sendo este, essencialmente texto, urls, imagens ou até vídeos.

A rede também cunhou o termo de "tweet" para designar uma publicação e "retweet" para o ato de partilhar o "tweet". Assim, esta rede social, tem como característica base a partilha de informação de forma rápida, casual e direta, permitindo que, posteriormente, esta seja partilhada de forma simples, pelo "retweet", ou adicionando um comentário à informação partilhada "quote tweet". Cada tweet pode, também, ser comentado e os utilizadores podem, ainda, demonstrar o seu apreço pela publicação utilizando o like.

Qualquer publicação nesta rede social tem como predefinição base ser pública e de fácil acesso a qualquer pessoa, seja esta utilizadora do Twitter ou alguém "de fora" consultando a plataforma. Contudo, o conteúdo partilhado por cada utilizador, apesar de visível a qualquer um, é apenas diretamente disseminado para os utilizadores que "seguem" o criador original. Esta característica da visibilidade das publicações, pode ainda ser ajustada, caso o utilizador assim o deseje.

Estas propriedades intrínsecas do Twitter são acompanhadas pela plataforma através do agrupamento de publicações em temas que por si criam as "tendências" dentro da plataforma. Para facilitar a disseminação da informação em torno de um tema, as "#" (hashtags) podem ser utilizadas para identificar uma dada publicação como relacionada com um certo tema.

3.2. API

3.2.1. Twitter API

De modo a disponibilizar os dados criados na rede Twitter, de uma maneira mais otimizada e padronizada, a plataforma disponibiliza uma API que permite que a utilização de todas as funcionalidades da rede social seja feita de forma "programática", como também, coletar a informação sobre utilizadores e o conteúdo que estes criam.

A API permite acesso aos dados através de diferentes níveis de permissões que os utilizadores podem ter de acordo com o plano de acesso. Para esta investigação, foi concedido o plano de "investigador académico", pela equipa do Twitter, à sua API. Este plano permite acesso aos dados em tempo real, bem como, aos dados históricos definidos como "públicos" e as suas características. Este plano permite o acesso a um máximo de dez milhões de tweets por mês e termos de pesquisa dos mesmos até 1024 caracteres.

Os dados da API são disponibilizados em formato JSON, distinguidos por objetos "Tweet" e "User" que posteriormente têm um conjunto de propriedades que podem ou não ser recolhidas conforme o plano do utilizador da API, e caso este assim o deseje. São ainda disponibilizados múltiplos URLs de conexão à API que permitem fazer diferentes tipos de recolha de dados (quando conectados por uma aplicação cliente), dividindo-se em URLs cujo objeto primário de retorno é o "tweet" e outros que visam direcionar a recolha de dados sobre o(s) utilizador(es) da rede. Cada ponto de recolha tem limites diferentes quanto à quantidade de informação a ser devolvida. Os pontos em torno de tweets são os que têm limites mais abrangentes em comparação com os de recolha sobre utilizadores que por norma são bastante limitados na quantidade de dados retornada por período de tempo.

3.2.2. Full archieve search

Para a recolha dos dados foi utilizado o endpoint tweets/search/all da API, que permite ter acesso a todos os tweets públicos existentes no arquivo de dados do Twitter, desde 2006 até ao presente, de acordo com os termos de pesquisa selecionados. Todos os tweets acessíveis por este endpoint são retornados pela ordem inversa à cronológica. Este endpoint permite realizar um pedido por segundo, retornando por defeito dez resultados por resposta até um máximo de 500, permitindo usar operadores de pesquisa exclusivos (são identificadores que permitem filtrar a pesquisa de acordo com a sua função, por

exemplo, o operador “from: twitterdev”, garante que a pesquisa apenas retorna tweets da conta twitterdev) para pesquisas com modo de acesso “académico”. De notar que o limite máximo de resultados por resposta pode ser reduzido, no caso da requisição de expansões dos objetos básicos de resposta como as “anotações de contexto”, que reduzem o máximo de 500 para 100 resultados por resposta.

3.2.3. Termos de pesquisa

A fim de conseguir tirar melhor proveito da API do Twitter, e, por conseguinte, dos seus *endpoints* de recolha de tweets, é necessário usar os termos de pesquisa corretos, para que os tweets devolvidos sejam próximos do pretendido. Deste modo, a API disponibiliza vários operadores de pesquisa e utilização de operadores booleanos. Operadores como o *is:retweet* ou *is:reply* permitem filtrar os tweets devolvidos para que estes sejam somente retweets e respostas a tweets, respetivamente. O operador “-” representa a negação e pode ser usado em combinação com outro operador para negar o seu efeito.

Relativamente à pesquisa de tweets por localização geográfica, o operador place_country filtra tweets cujo código do país associado seja igual ao usado no filtro, no formato ISO alfa-2. Para a filtragem de tweets de uma determinada língua, existe o operador lang, que retorna tweets que tenham sido previamente classificados pelo Twitter como pertencentes à língua especificada. Esta especificação tem de corresponder a uma língua do sistema BCP 47. De momento, esta classificação apenas denomina um tweet como sendo de uma língua apenas, mesmo que utilize palavras de mais do que uma.

3.3. Processo de recolha

3.3.1. Extração de tweets e respetivos utilizadores

O processo de extração de dados usou a API do Twitter como ponto de coleta de dados. Este processo e projeto foi desenvolvido em Python, e recorreu à biblioteca *tweepy* como ponto intermédio entre o programa e a API, de modo a simplificar o processo de estabelecimento de autenticação, conexão, agregação e formatação dos dados enviados pelo Twitter. Esta biblioteca também simplifica a criação do pedido à API, a definição dos filtros a aplicar e as propriedades de expansão a retornar para os tipos de objetos possíveis de recolher.

De forma a construir uma amostra de dados representativa do Twitter tendo em conta o objetivo da investigação e que tornasse possível reduzir ao máximo a existência de viés[1] nos dados, a recolha define como objetivo obter, de uma perspetiva macro, tweets de todos os dias desde 01-01-2019 a 31-12-2021, totalizando 1.603.659 tweets ao longo deste período com em média, 227.982 tweets por dia da semana, de forma que o resultado final consiga construir uma amostra do comportamento da informação durante os três anos.

$$lang{:}\,en\ place\_country{:}\,US\ -url\ -is{:}\,retweet\ -is{:}\,reply\ -the\ the$$

Figura 2: Termos de pesquisa usados para recolha dos tweets

Para dar resposta à recolha de dados proposta, procurou-se que todos os tweets fossem recolhidos em inglês e com origem nos Estados Unidos da América, de modo a garantir ao máximo, a coerência no conteúdo dos mesmos e que a língua predominante fosse o inglês. Por outro lado, foram ignorados todos os retweets, e respostas a tweets. De modo a satisfazer a condição da API de que o critério de pesquisa não pode conter somente operadores, e tendo em conta que a investigação não pretendia recolher tweets apenas de um determinado assunto/tema, foi feita uma pesquisa pelo termo "the" que é amplamente usado na língua inglesa permitindo captar uma maior variedade de tweets.

[1] Tendência de algo. Distorção na maneira de observar, de julgar ou de agir.

3.3.2. Metodologia de recolha de dados

Tendo em conta as características da API do Twitter, em particular as da recolha de tweets do *endpoint tweets/search/all*, isto é, do ponto que permite recolher tweets desde a criação da rede social em 2006 até aos dias de hoje, esta recolha devolve sempre os tweets de forma decrescente, a contar do período máximo definido até à data inicial que se pretende. Isto é, por exemplo, para recolher tweets de um dia em particular, definido pelos limites 01-01-2020 a 02-01-2020, há aproximadamente 70 milhões de tweets para recolher, e sendo que o objetivo é recolher uma pequena amostra desse dia, o resultado traduz-se na recolha de tweets de uma certa altura desse dia, pois o número de tweets publicados nesse espaço de tempo, são superiores aos que se pretendem recolher. Isto levanta um problema, uma vez que pode originar conclusões tendenciosas devido à falta de representatividade de tweets originados noutras fases do dia.

Para dar resposta a esta restrição, é feito um processo de recolha de tweets, onde para além de recolher tweets de todos os dias dos três anos, divide a recolha dos tweets de cada dia em cinco fases, noite cerrada (01:00:00), manhã (08:00:00), depois de almoço (14:00:00), fim de tarde (18:00:00) e noite (22:00:00). Esta recolha faseada permite dar ao conjunto de dados recolhidos, uma representação das diferentes fases do dia, permitindo que seja possível compreender os comportamentos da informação em diferentes períodos, do ano, semanas, dias da semana e fases do dia. Deste modo, foram feitos 5475 pedidos de dados à API, que corresponde a 5 pedidos (um para cada fase do dia) por dia para os 1095 dias dos três anos. Este processo leva entre a seis a oito horas continuas para coletar dados, uma vez que o processo corre de forma síncrona.

Esta metodologia de recolha também insere uma restrição na amostra dos dados recolhidos, sendo ela, a falta de perceção do volume total de tweets para um dado dia ou até fase do dia, pois nesta metodologia, é definido um volume artificial de tweets por período de recolha. Apesar disso, e uma vez que o foco da investigação, não é analisar a evolução de um tema em particular, a existência desta restrição é negligenciada pela pouca relevância para o estudo da popularidade da informação, de uma forma geral.

3.3.3. Propriedades dos dados recolhidos

De forma a recolher os dados necessários para realizar a análise pretendida, todos os tweets foram recolhidos recorrendo aos campos “author_id”, “created_at”, “conversation_id”, “referenced_tweets”, “public_metrics”, “source”, “reply_settings”, “lang”, “context_annotations”. De modo a obter os dados relativos aos utilizadores que originaram cada tweet, a recolha foi feita usando a expansão “author_id”, que permite retornar para cada tweet o objeto “User” do seu criador. Este objeto por sua vez foi recolhido com os campos “created_at”, “verified” e “public_metrics”.

Decompondo o significado de cada um dos campos extra requisitados para os objetos *Tweet* e *User* temos o seguinte esquema:

Tweet:

- author_id, indica o id do utilizador que originou o tweet
- created_at, indica a data em que o tweet foi originado no formato ISO 8601 (<AAAA-MM-DD>T<HH:MM:SS>Z)
- conversation_id, indica o ID do tweet original de uma conversa a que um dado tweet pertence
- referenced_tweets, devolve a lista de tweets a que um dado tweet se refere, indicando o tipo de tweet (retweet, quoted)
- public_metrics, devolve a métricas sobre um dado tweet à data da recolha (número de retweets, likes, quotes e comentários),
- source, indica o tipo de dispositivo usado para publicar o tweet (exemplo: “Android” ou “iPhone”),
- reply_settings, indica quem pode responder a um dado tweet (“todos”, “utilizadores mencionados” ou “seguidores”),
- lang, indica qual a linguagem identificada em que o tweet foi escrito, por fim o campo
- context_annotations, devolve informações sobre uma análise de tópico feita ao conteúdo do tweet.

User:

- created_at, indica a data, no formato isso 8601, de criação da conta do utilizador que originou o tweet,

- verified, designa se a conta do utilizador é de interesse público e é verificada[2] pelo Twitter confirmando que a conta é autêntica,
- public_metrics, devolve detalhes sobre o potencial alcance da conta e a sua atividade (número de seguidores, número de seguidos, número de tweets).

3.3.4. Anotações de contexto e análise de tópico

Com a finalidade de elaborar uma análise de conteúdo dos tweets recolhidos, todos eles foram coletados com as respetivas anotações de contexto. As anotações de contexto, são um campo sobre o tweet que devolve uma análise previamente feita pelo Twitter sobre o possível assunto/tópico na qual se insere um dado tweet. Contudo, apesar de todos os tweets serem previamente analisados neste sentido, devido ao conteúdo dos mesmos, nem sempre é possível decifrar o seu tema.

As anotações de contexto dividem-se em duas categorias, a mais especifica denominada por entity que define dentro do assunto o tema real, como por exemplo a pessoa ou local sobre o qual o tweet fala. Estas são diretamente contextualizadas do conteúdo do tweet e acompanhadas do local no texto onde se inserem.

Por outro lado, existe a categoria mais abrangente denominada por "context" que é derivado de uma análise de tópico ao texto do tweet. Um tweet pode ser classificado como pertencente a um ou mais tópicos, onde estes são atribuídos por uma lista com mais de 50 domínios, disponibilizados pelo Twitter.

3 - TV Shows	46 - Brand Category	79 - Video Game Hardware	115 - Video Game Conference
4 - TV Episodes	47 - Brand	84 - Book Music Genre	116 - Video Game Tournament
6 - Sports Events	48 - Product	85 - Book Genre	117 - Movie Festival
10 - Person	49 - Product Version	86 - Movie	118 - Award Show
11 - Sport	54 - Musician	87 - Movie Genre	119 - Holiday
12 - Sports Team	55 - 56 - Actor	88 - Political Body	120 - Digital Creator
26 - Sports League	58 - Entertainment Personality	89 - Music Album	122 - Fictional Character
27- American Football Game	60 - Athlete	90 - Radio Station	130 - Multimedia Franchise
28 - NFL Football Game	65 - Interests and Hobbies Vertical	91 - Podcast	132 - Song
35 - Politicians	66 - Interests and Hobbies Category	92 - Sports Personality	136 - Video Game Personality
38 - Political Race	67 - Interests and Hobbies	93 - Coach	137 - eSports Team
39 - Basketball Game	68 - Hockey Game	94 - Journalist	138 - eSports Player
40 - Sports Series	71 - Video Game	110 - Viral Accounts	139 - Fan Community
45 - Brand Vertical	78 - Video Game Publisher	114 - Concert	

Figura 3: Lista dos 50+ domínios usados para classificação do tema de um tweet

[2] Conta identificada com um selo azul, atribuído pelo Twitter, que informa que uma é uma conta de interesse público.

Como sendo uma propriedade extra de um objeto *tweet* recolhido, estas anotações de contexto assumem a seguinte forma:

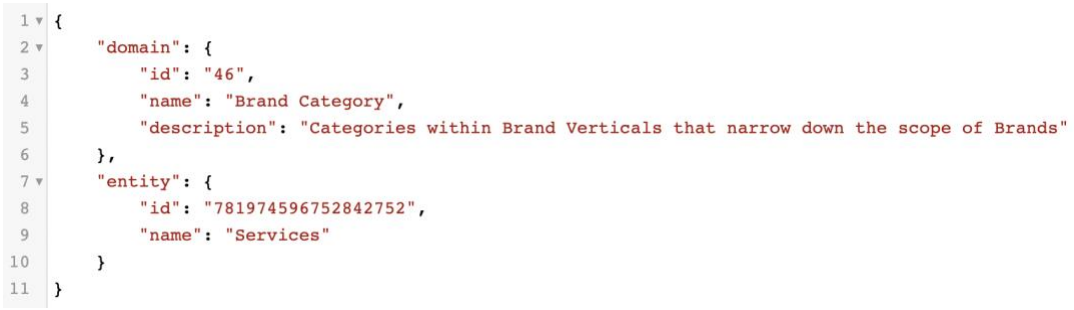

Figura 4: Objecto de anotação de contexto de um tweet em formato JSON

Devido às restrições nos operadores de pesquisa de *tweets* existentes, o operador *context* apenas pode ser usado com o formato *context: domain_id.entity_id* e dado que, à data, o Twitter não providencia uma lista com os *ids* das entidades, não é possível recolher apenas *tweets* em que foi possível contruir o objeto das anotações de contexto dos *tweet*, ou seja, tweets em que o Twitter disponibiliza a quando da sua recolha o tópico do mesmo. Assim, para os dados recolhido apenas aproximadamente, 35% dos *tweets* recolhidos têm informação sobre o seu tópico.

3.3.5. Retweets e alcance de um tweet

A fim de colmatar a falta de informação quanto ao verdadeiro alcance de um tweet, e uma vez que saber o número de retweets e quote tweets, ainda que bons indicadores da popularidade de um tweet, não transmite informação sobre quantas pessoas poderão ter tido contacto com o tweet (essa métrica é exclusiva a utilizadores da API do Twitter em modo de *PowerTrack*, que é pago). Assim sendo, adotamos a seguinte estratégia.

Sabendo que um tweet quando é partilhado por um utilizador, tem o potencial de aparecer no feed de todos os seus seguidores, pode-se assumir que o potencial de alcance de um tweet é igual ao número de seguidores do utilizador que criou o tweet. Sempre que um tweet é partilhado, este tem o potencial de ser visto por todos os seguidores do utilizador que partilhou o tweet original, e o mesmo se verifica para utilizadores que façam quote do tweet original. Assim, o potencial de alcance de um tweet pode ser definido por:

Equação 1: Formúla usada para calculo do alcance de um tweet

$$Total\ alcance = Nº\ seguidores\ "originais" + \sum Nº\ seguidores\ de\ cada\ retweeter + \sum Nº\ seguidores\ de\ cada\ quoter$$

De maneira a obter uma estimativa deste valor, é necessário saber o número de seguidores de cada utilizador que partilhou ou quoted um tweet. A API do Twitter disponibiliza o endpoint /2/tweets/:id/retweeted_by e /2/tweets/:id/quote_tweets, mas estes endpoints são muito limitados na quantidade de dados que se pode recolher, permitindo apenas 75 pedidos a cada 15 minutos, ou seja, por cada 15 minutos apenas se consegue recolher os retweeters e quoters de 75 tweets.

Dado que recolher os retweeters e quoters de todos os tweets que continham informação sobre o tópico do tweet e foram partilhados, teriam que ser recolhidos os dados para 64.878 tweets, o que para os endpoints existentes, tornava o processo inviável. Assim, foi feita uma recolha utilizando o endpoint tweets/search/all utilizando como método de pesquisa a seguinte forma:

$$(url: Twet\_id)\ (is: retweet\ OR\ is: quote)$$

Figura 5: Termos de pesquisa usados para recolher os retweeters

Dada o limite de 1024 caracteres para cada pesquisa, a componente identificadora de cada tweet (url:tweet_id) era repetida quantas vezes coubesse dentro do limite de caracteres, permitindo, com apenas um pedido à API, recolher os dados de em média 36 tweets, visto que o tamanho do tweet_id varia.

$$(url: tweet\_id_1\ OR\ url: tweet\_id_2)\ (is: retweet\ OR\ is: quote)$$

Figura 6: Termos de pesquisa usados para recolher os retweeters para mais que um tweet de cada vez

Este método de pesquisa retorna todos os retweeters e quoters associados a um dado tweet. Contudo, é de notar que em certos casos nem sempre a totalidade dos retweeters era devolvida. Para além disso, foi ainda notado que por vezes o número de retweets ou quotes que um tweet tem não corresponde à realidade, pois como os dados são contados à data da recolha estes não deixam de ser mutáveis.

Após recolher os retweeters e quoters de cada tweet selecionado, é então possível satisfazer as condições de forma a calcular o número aproximado do potencial de utilizadores alcançados por cada tweet. Este dado é importante porque, por exemplo, para dois tweets cada um com um retweet, um partilhado por um utilizador com 50 seguidores não terá a mesma probabilidade de se tornar viral do que um partilhado por um utilizador com 5.000 seguidores, ainda que ambos tenham o mesmo número de retweets.

3.3.6. Sentimento de um tweet

De modo a complementar a análise de tópico, já oferecida pelo API do Twitter, foi feita uma análise de sentimento ao conteúdo de cada tweet, isto é, o texto de cada tweet foi classificado como podendo ser "Positivo", "Negativo" ou "Neutro". Para alcançar estas classificações foi usada uma biblioteca de python que complemente o modelo de análise de texto, VADER. Este modelo pode ser usado não só para analisar a polaridade de um texto, mas também a sua intensidade. Para efeitos de simplificação do processo de análise de resultados, foram apenas considerados os resultados da polaridade (classificação atrás citada), de cada tweet como base da análise de sentimento.

3.3.7. Outliers

De modo a evitar possíveis discrepâncias nos resultados e análises, fez-se uma remoção de valores que se encontravam muito distantes da maioria dos seus pares. Assim, fez-se uma análise inicial da distribuição das variáveis recorrendo a histogramas e caixas de bigodes (boxplot) para variáveis métricas e gráficos de barras para variáveis não-numéricas. Com uma melhor compreensão do modo de distribuição das variáveis, foi feita uma remoção dos valores que mais impacto negativo poderiam trazer. Esta remoção foi feita sempre mantendo a regra, de não remover mais de 3% da amostra de dados.

$$followers < 100000 \,\&\, following < 30000 \,\&\, retweet\_count < 1000 \,\&\, like\_count < 400 \,\&\, seniority < 17$$

Figura 7: Critérios usados para remover outliers dos dados

Para compreender a remoção feita é preciso ter em conta que os dados apresentavam grandes disparidades. As principais características mostravam grandes níveis de

"outliers" e analisando a natureza destes dados, a sua remoção não era direta. Por exemplo, removendo todos os valores que se desviam para além dos limites superiores das caixas de bigodes, eliminaria por consequência, uma percentagem significativa da amostra. Este impacto é compreensível, quando se tem em conta uma possível correlação entre o número retweets de um tweet e o número de seguidores que a conta de origem tem. Assim remover contas com mais de 5000 seguidores (onde começam os outliers) representaria a remoção de 23% dos posts com retweets. Isto tem especial impacto quando, em média, num dado ano, o número de posts recolhidos com retweets (> 0) é de 20%. O mesmo se verifica na perspetiva contrária, em que remover tweets com retweets acima de 5 retweets, que representam 2,5% da amostra, significava remover aproximadamente 15% dos tweets com retweets.

Deste modo os valores limites, arbitrariamente definidos, foram escolhidos por serem os que permitiam a maior remoção possível de outliers sem exceder a remoção de 3% dos dados e apenas removendo 2,97% dos tweets com retweets.

3.4. Tratamento de dados

Uma vez que a recolha de dados sobre os tweets e seus utilizadores é feita através da API oficial do Twitter, a maioria deles já se encontra diretamente em formato útil para processamento. Contudo, de forma a personalizar os dados para o objetivo da investigação, algum tratamento é requerido. Assim, este processo de tratamento dos dados recolhidos separa-se em dez partes distintas, aplicadas sequencialmente para obter o resultado final.

1ª Fase: **Carregamento dos dados**, onde os dados, dos tweets e suas respetivas contas de origem, são carregados em memória e ordenados por ordem temporal, do mais antigo para o mais recente. Para os dados das contas de utilizadores, são removidas contas duplicadas. Esta duplicação pode acontecer pois, uma vez que, os utilizadores são recolhidos por adjacência a cada tweet recolhido, podem ocorrer casos em que mais do que um tweet do mesmo utilizador é recolhido, levando a duplicação da informação.

2ª Fase: **Cálculo do alcance dos tweets**, de forma a calcular uma aproximação do alcance de um tweet, é feita uma remoção dos utilizadores (retweeters) que estavam duplicados e, em seguida, procede-se a uma junção dos dados dos retweets com os respetivos retweeters. Uma vez que, durante a recolha dos retweeters, e devido à metodologia usada (em parte recorrendo ao *URL* do tweet original), alguns tweets originais foram recolhidos para além dos retweets e quote tweets pretendidos, uma vez que continham no texto do tweet, o url do tweet original. Contudo, visto que tweets que apenas mencionam o *URL* do tweet original não contam para a métrica do número de retweets de um tweet, estes casos foram excluídos. O exemplo em baixo, ilustra um tweet recolhido mencionando o *URL* do tweet para o qual se pretendia recolher os respetivos retweets.

$$\#FlyEaglesFly\ \#PhillysvsEveryone\ https://t.co/G9H4EnmKgq\ https://t.co/eXbEfo7wCm$$

Figura 8: Exemplo de menção de tweet removido que incluí URL do tweet original que se procurava

Seguidamente, para cada *tweet* é feita a extração do *id* do *tweet* referenciado, mas também para cada retweet recolhido, de forma a transformar o objeto devolvido pela API num campo mais fácil de processar.

$$[ReferencedTweet(id: 1212175890785996801, \text{type: retweeted})]$$
$$\rightarrow 1212175890785996801$$

Figura 9: Exemplo da transformação aplicada de forma a extrar o id do tweet "pai"

Por fim, para cada tweet original onde tenha sido possível identificar o seu tópico, e com pelo menos um retweet, é obtido o número aproximado do máximo de utilizadores alcançados, somando o número de seguidores dos utilizadores que fizeram todos os retweets que referenciavam o *id* do *tweet* original.

3ª Fase: **Processamento dos tópicos**, que se inicia por um pré-processamento dos tópicos e dos seus respetivos IDs onde os tweets sem tópicos são ignorados e apenas o tópico mais provável da lista de opções é retido para efeitos de classificação do tema do tweet. Por fim, os tópicos são agrupados para uma categoria mais abrangente segundo o método descrito anteriormente, na secção 3.3.3.

4ª Fase: **Classificação de sentimento**, utiliza a ferramenta VADER para classificar o texto de cada tweet como sendo "Positivo", "Neutro" ou "Negativo".

5ª Fase: **Identificação de hashtags**, onde se faz um reconhecido e marcação de todos os tweets que contêm ou não hashtags no corpo de texto.

6ª Fase: **Popularidade**, a fim de conseguir tirar proveito de algoritmos de *machine learning* com base em aprendizagem supervisionada, foi também feita uma classificação binária prévia da popularidade de cada *tweet*. Assim, foi considerado que um tweet poderia ser "Popular" sempre que este tivesse pelo menos um retweet ou "Não popular" quando a regra anteriormente mencionada não fosse cumprida. Estas duas classificações foram caracterizadas por "1" para os *tweets* populares e "0" para os *tweets* que não o são. O critério para a definição de tweet popular começa num número mínimo de partilhas muito baixo, uma vez que a maioria dos *tweets* recolhidos não tem sequer um *retweet*, ou seja, mesmo simplificando o significado de popularidade, a divisão dos dados resultante continua ainda com um desequilíbrio considerável na representação de *tweets* populares, onde apenas 16% foram considerados como populares.

7ª Fase: **Fundir amostras**, resultando num *dataset* final convergindo os dados de cada tweet com os dados do utilizador que os originaram, para facilitar as análises futuras, ou seja, permitindo ter todas as informações sobre um *tweet* no mesmo sítio.

8ª Fase: **Transformação de variáveis**, nesta fase a variável *timestamp* é decomposta em variáveis como mês, ano, etc.… e as variáveis categóricas são transformadas em numéricas através da aplicação do método *LabelEncoder*, este processo permite facilitar o processamento destas variáveis durante a fase de análise. Para além disso, é ainda também aplicada a transformação às variáveis categóricas recorrendo ao *OneHotEncoder*, com o intuito de normalizar estas variáveis numa escala entre [0,1], de forma a obter um formato mais acessível para os modelos de aprendizagem automática. Ambos os métodos mencionados são utilizados recorrendo à biblioteca de *python sklearn*(*Scikit-Learn*, n.d.). Por fim, é calculada a antiguidade de todas as contas dos utilizadores recolhidos à data da publicação de cada tweet, bem como para todos os retweeters.

9ª Fase: **Remoção de outliers,** de forma a remover valores que possa criar tendências nos dados, é aplicado o filtro manual, que remove até 3% da amostra de dados, explicado na secção 3.3.6.

10ª Fase: **Guardar dados,** para finalizar, dos quatro subconjuntos de dados recolhidos, de um dado ano (tweets, utilizadores, retweets e utilizadores dos retweets), obtêm-se dois novos conjuntos de dados totalmente tratados e processados que são guardados em ficheiros no formato *csv* para uso na fase de análise e previsão. A escolha de armazenamento dos dados em formato *csv* foi de forma a simplificar o processo de processamento dos dados, uma vez que o acesso aos mesmos não requeria ser num formato regular, mas apenas para quando necessário proceder a uma análise, momento no qual os dados são carregados em memória, com recurso a bibliotecas de *python* como *pandas*.

3.5. Processo de análise e previsão

Após os dados serem processados e tratados, acabando todos com o mesmo formato, estes são fornecidos à componente do sistema responsável pela produção automática das análises a realizar. A fase de análise dos dados divide-se em duas partes principais: (1) a análise dos dados sobre os tweets em si e a (2) análise dos dados sobre os utilizadores que originaram os respetivos tweets bem como sobre os utilizadores que partilharam os tweets "populares". Durante este processo, uma série de análises são definidas e aplicadas de modo a sintetizar o melhor possível a informação recolhida para fundamentar a investigação, sendo que cada análise gera um resultado sob a forma de tabela e um gráfico que faz a visualização dos dados de forma a facilitar a sua interpretação. Os resultados destas análises são apresentados e discutidos no capítulo seguinte.

Depois da análise dos dados, segue-se a parte final do programa que tenta usar os dados recolhidos para previsão do comportamento e popularidade futura de um tweet de um certo tópico à escolha. Para isso, os dados recolhidos são agrupados e divididos em duas componentes: dados de treino, onde o modelo irá "aprender" e dados de teste que serão usados para validação do desempenho do modelo criado. Os resultados do modelo são apresentados através de um gráfico ilustrativo gerado automaticamente, bem como uma tabela resumo com as principais métricas de avaliação, processo este que é caracterizado no capítulo seguinte

Todo o código desenvolvido que permite a recolha, análise, previsão e os respetivos dados recolhidos, podem ser consultados no *Github* de apoio ao projeto em https://github.com/MiguelDordio/InformationFlow.

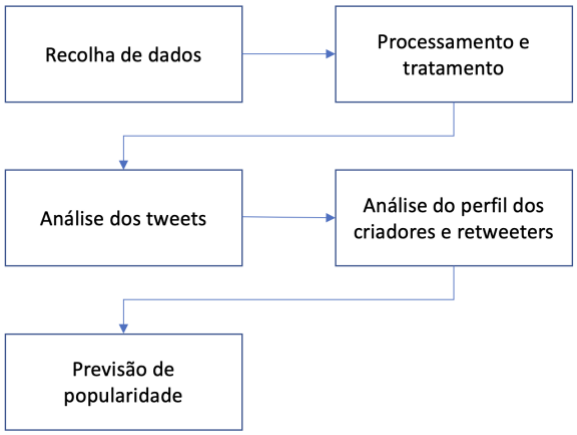

Figura 10: Diagrama do fluxo dos dados criado pelo programa desenvolvido

Capítulo 4 – Análise e discussão dos resultados

4.1. A importância do espaço temporal na popularidade de um tweet

De forma a analisar a importância do espaço temporal (fase do dia) na popularidade de um tweet, foram analisados tweets publicados em diferentes fases do dia e ao longo dos diferentes dias da semana. Assim, estes foram recolhidos às 01:00:00 (middle of the night), 08:00:00 (morning), 14:00:00 (afternoon), 18:00:00 (dusk) e 22:00:00 (night).

Como é visível pelo gráfico da figura 11, o período da manhã é consistentemente a fase do dia em que existe, em média, que regista menos número de partilhas e gostos. Contudo a hora de almoço, é fase do dia onde, em média, mais retweets são registados, com uma tendência a aumentar ao longo dos anos. Verifica-se também, que tanto o período do fim de tarde como o da noite mostram manter o nível de atividade iniciado à hora de almoço, culminando num pequeno decréscimo com o aproximar das horas mais tardias da noite. Este comportamento é muito similar para a média de "gostos" registada, uma vez que segue a mesma tendência observando-se, no entanto, valores médios mais elevados em relação aos retweets. Para além disso, uma outra diferença, é o pico de atividade que é alcançado durante a noite em comparação com o pico de retweets, que ocorre à hora de almoço. Verifica-se também um crescente uso de gostos, com a média a aumentar exponencialmente para o ano de 2021, nas diferentes fases do dia.

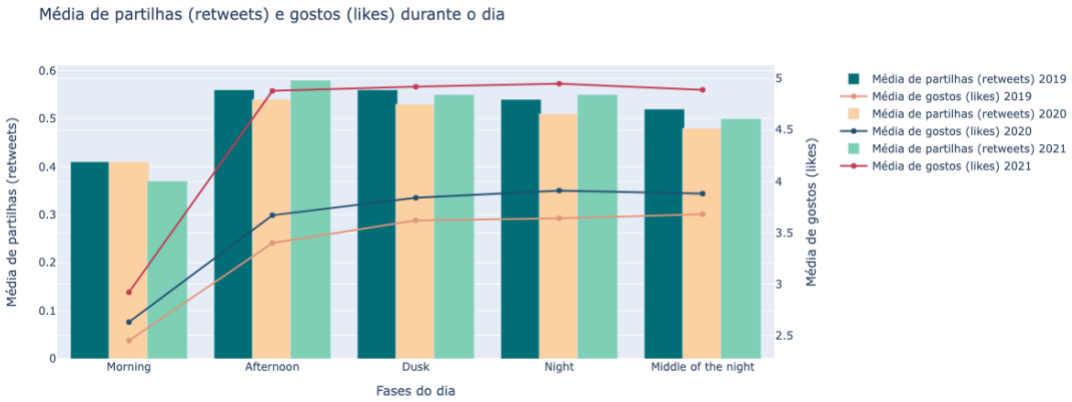

Figura 11: Média de partilhas e gostos durante o dia entre 2019 e 2021

Quanto aos dias da semana, o dia que registava mais atividade de retweets em 2019 foi quinta-feira e, com o passar dos anos, a tendência verificada é um aumento da atividade mais voltada para o fim da semana onde em 2021 o pico já é à sexta-feira.

Apesar disso, de um modo geral, os níveis de atividade média de retweets permanecem sem grandes variações durante o decorrer da semana e mantêm o comportamento ao longo dos anos. O mesmo se verifica para o número médio de "gostos", que, tal como para os dados das fases do dia, apresentam um valor médio ligeiramente mais elevado que os retweets e onde o pico de atividade acontece aos sábados, mantendo uma tendência crescente na sua quantidade média ao longo dos anos. Contudo, antes deste crescimento, o número médio de "gostos" caí significativamente em 2020, da mesma maneira que o número médio de partilhas também decresce nesse ano. Este fenómeno pode ser derivado pela ocorrência e aparição do covid.

Assim sendo, verifica-se que o número médio de retweets apresenta uma fraca. ou até, ligeiramente, moderada correlação com o número médio de gostos, quando tendo em conta as fases do dia, isto verifica-se quando é aplicada a correlação de *Spearman* que devolve um valor de $\rho = 0{,}39$ entre as duas variáveis. O método de *Spearman* foi escolhido uma vez que é uma medida de correlação não paramétrica e as variáveis em causa não seguem uma distribuição normal. Tendo também em conta a força da ligação entre as duas variáveis e recorrendo ao método de *Kendall*, também se confirma um valor baixo de $\tau = 0{,}20$, ainda assim, a correlação demonstra ser positiva entre as duas variáveis, ou seja, quando uma aumenta a outra por norma também o faz. Por outro lado, quando tentamos identificar a melhor altura para publicar um tweet, a fase do dia em que este é feito demonstra ter maior relevância no desempenho da sua disseminação, em relação ao dia da semana, com a hora de almoço a demonstrar ser a melhor aposta. Apesar do dia da semana não evidenciar ter tanto significado, parece existir uma tendência para sexta-feira ser o melhor dia da semana para publicar um tweet.

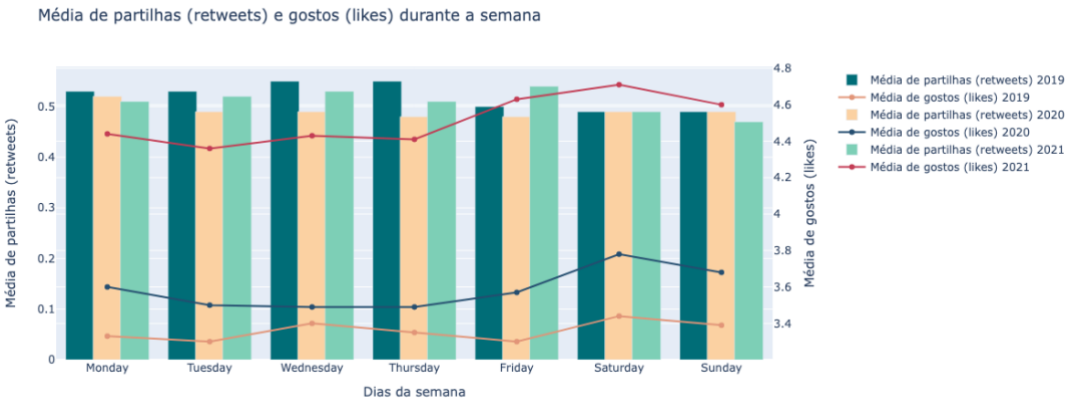

Figura 12: Média de partilhas e gostos durante a semana entre 2019 e 2021

4.2. Importância do sentimento do conteúdo no desempenho dos tweets

Uma vez que muito do sucesso do desempenho da popularidade de um tweet provêm do sentimento que este transmite, foram feitas análises às diferentes categorias de sentimento que um tweet pode transmitir: positivo, neutro e negativo (categorias obtidas através da biblioteca VADER durante o processo de tratamento dos dados, referido no capítulo 3), mas também, o efeito de disseminação que cada uma conseguiu alcançar durante o período de tempo, entre 2019 e 2021.

A figura 13 mostra que tweets com sentimento positivo conseguem um maior número médio de partilhas, onde é apresentada uma ligeira vantagem em relação a tweets com sentimento negativo. Já os tweets com sentimento neutro, apesar de demonstrarem ser os que, em média, menos partilhas e likes conseguem, não se encontram muito longe das outras categorias. Este fenómeno apresenta o mesmo comportamento para o número médio de gostos obtidos nos tweets, ou seja, tweets com sentimento positivo não só conseguem obter mais partilhas como também mais gostos. Já tweets com sentimento neutro demonstram obter, em média, menos gostos e partilhas que qualquer outro sentimento.

Estes comportamentos apresentam consistência ao longo dos anos, não só para o número médio de retweets como para o número médio de "gostos". Para além disso, ambas as métricas desta vez demonstram uma correlação pelo método de *Spearman* mais forte, de $\rho = 0{,}67$, ou seja, quando o sentimento dos tweets é tido em conta, o número de partilhas e gostos apresenta uma correlação forte e positiva entre estas variáveis, apresentando a mesma tendência, para as mesmas categorias. De destacar também o contínuo aumento exponencial do número médio de gostos, para os mesmos valores médios de retweets ao longo dos anos.

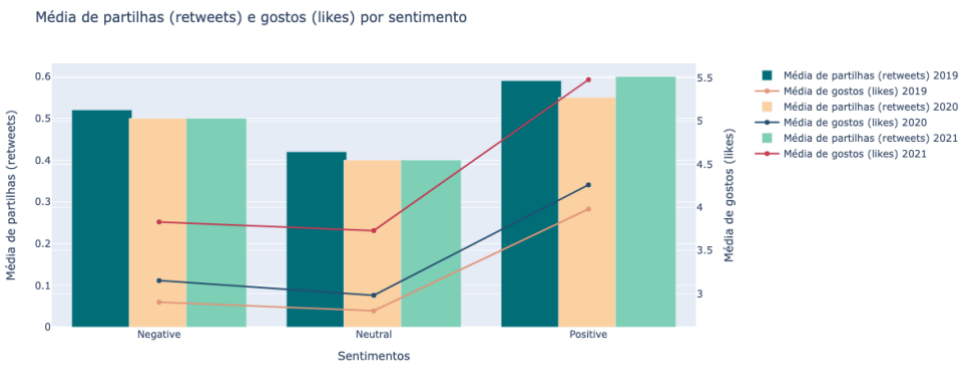

Figura 13: Média de partilhas e gostos por sentimento dos tweets expresso entre 2019 e 2021

4.3. Tweets contextualizando os seus tópicos

4.3.1. Análise da disseminação dos tweets por tópicos

Ainda com hipótese de que o conteúdo de uma publicação é um dos fatores que melhor indica a quantidade de pessoas que o poderão ver, e, por conseguinte, o nível de interação que irá ter, e de forma a ganhar uma melhor compreensão e capacidade para quantificar o impacto desta característica, foi feita uma análise aos 14 tópicos previamente definidos como classificadores do conteúdo de um tweet procurando analisar o desempenho de cada um nas diferentes circunstâncias. Para que os resultados dos diferentes tópicos pudessem mais facilmente ser comparados entre si, uma vez que as amostras recolhidas de tweets para cada tópico são de dimensões consideravelmente diferentes, e de modo a facilitar a sua análise, foi usada como medida a percentagem de tweets com partilhas de cada um dos tópicos analisados, ou seja, esta percentagem é obtida da seguinte forma:

Equação 2: Formúla para calcula da percentagem de partilhas para um dado grupo de tweets

$$Percentagem = (nº\ de\ tweets\ com\ partilhas / nº\ total\ de\ tweets) \times 100$$

Deste modo, e observando a figura 14, verifica-se que para todos os tweets publicados sobre cada tópico, em geral, a percentagem dos quais conseguem obter pelo menos uma partilha apenas varia entre 14% a 27%, apresentando a maioria dos tópicos valores mais próximos do limite inferior registado, ou seja, menos de $\frac{1}{4}$ dos tweets consegue obter partilhas. Apesar disso, certos tópicos demonstram conseguir melhores resultados que os seus pares, como os tópicos "noticias", "música" e "desporto" que são os que apresentam valores de maior tração nas redes sociais, sendo os que, de forma consistente, se conseguem destacar dos restantes.

Por outro lado, a percentagem de tweets que consegue obter "gostos" demonstra ser superior aos que conseguem retweets, atingindo valores entre os 31% e 100%, sendo que a maioria dos tópicos regista valores entre os 45% e os 65%. Contudo, em relação aos gostos, os tópicos têm maior percentagem de gostos nos seus tweets, são "notícias" e "desporto", sendo o tópico "politica" o que menos likes consegue. Assim, verifica-se que os tópicos "notícias" e "desporto", se destacam tanto pela percentagem de retweets como gostos.

O tópico "livros", em 2019, apresenta uma percentagem de tweets com partilhas consideravelmente alta, destacando-se dos restantes, contudo, este "pico" pode ser

explicado por alguns outliers que não foram removidos, pelo filtro aplicado e explicado no capítulo anterior. Posto isto, terá ganho maior destaque tendo em conta que a amostra de tweets sobre "livros" recolhidos em 2019, foi de apenas 208, comparado com tópicos como "pessoas" onde foi possível recolher uma amostra de 36.681 tweets só para 2019. O mesmo fenómeno pode explicar o "pico" para tweets sobre "férias" em 2019, contudo, para este caso a amostra recolhida foi ainda menor, de apenas um. Uma outra discrepância registada na análise foi a falta de dados sobre tweets com retweets em 2019 sobre "férias" e "notícias" o que não permitiu tirar conclusões sobre os seus desempenhos.

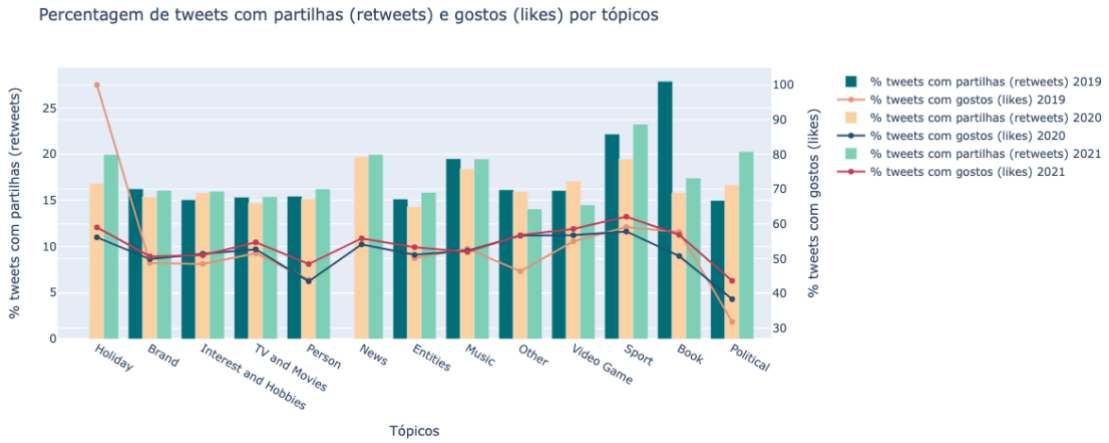

Figura 14: Percentagem de partilhas e gostos por tópico entre 2019 e 2021

Para além desta análise de popularidade, a figura 15, demonstra que a importância da fase do dia apresenta um impacto nos tópicos que conseguem maior disseminação ao longo do dia, sendo a hora de almoço a fase do dia, onde o desempenho dos tópicos se distingue das restantes, algo que se verifica acontecer durante os diferentes anos. Apesar disso, é notável a homogeneidade que existe entre as fases do dia e os diferentes tópicos abordados em cada uma delas, pois de forma geral, todos os tópicos, ao longo dos três anos, apresentam uma distribuição similar naquilo que é a percentagem de tweets que conseguem de facto serem partilhados.

Figura 15: Percentagem de partilhas por tópico durante o dia entre 2019 e 2021

Analisando em seguida o impacto do dia da semana nos diferentes tópicos, a figura 16, permite observar a igual distribuição dos tópicos ao longo da semana. Apesar disso, é de notar que, certos tópicos que de uma forma geral não conseguem um melhor desempenho, podem diferenciar-se quando tido em conta o dia da semana. Por exemplo, o tópico "interesses e hobbies" que consegue ser um tópico com maior atividade, quando tweets do mesmo são criados durante a semana, por exemplo à quarta-feira.

Figura 16: Percentagem de partilhas por tópico durante a semana entre 2019 e 2021

Recorrendo ao cálculo do alcance de um tweet definido no capítulo 3, secção 3.3.4, é possível verificar pela figura 17, que, em média, o alcance dos tweets dos tópicos analisados (do número de pessoas que, em média, recebe o tweet que aparece na sua linha temporal da rede social) não vai além das 5000 pessoas. Contudo, certos tópicos demonstram conseguir disseminar mais o seu conteúdo, como é o caso de tweets sobre "notícias" ou "pessoas", já tweets sobre "jogos digitais" demonstram ser os que menos pessoas alcançam. Também se verifica que ao longo dos anos o alcance aparenta ser consistente mantendo as já aferidas, diferenças entre os tópicos.

Figura 17: Alcance médio dos tweets por tópico entre 2019 e 2021

4.3.2. Análise do desempenho dos tweets por tópicos e sentimento

De maneira a aprofundar melhor a análise e o impacto que o conteúdo de um tweet tem na sua disseminação, foi feita uma análise onde se pretende dissecar, para cada tópico, o modo como cada tipo de sentimento que o tweet pode ter, e se este, impacta o seu sucesso.

Assim sendo, pela figura 18, é possível verificar tendências, em que para certos tópicos, como por exemplo, o "desporto" em 2019, tweets neutros conseguiam melhores resultados, mas que ao longo dos anos a situação se alterou, favorecendo em 2021 tweets com sentimento positivo. Por outro lado, tópicos como "música" e "marcas" permaneceram consistentes ao longo dos anos, favorecendo sempre tweets mais positivos.

Deste modo, os dados ilustram que, dependendo do tópico, certos tipos de sentimentos conseguem alcançar melhores resultados, sendo um fator a ter em conta,

e onde se verifica uma certa consistência ao longo tempo, ou seja, um tópico que demonstra ter um certo tipo de sentimento mais popular, tende a continuar a favorecer esse sentimento ao longo do tempo. Porém, nem sempre acontece, e dependendo das circunstâncias, que vão mudando com o passar dos anos, a polaridade entre preferências para certos sentimentos pode inverter-se para alguns tópicos, como por exemplo, é o caso dos tópicos "desporto" e "pessoas", onde se nota que o sentimento que melhor resultado consegue, deixou de ser o "negativo" para "positivo".

Figura 18: Percentagem de partilhas por tópico e sentimento 2019 e 2021

4.3.3. Análise do impacto da presença de hashtags no desempenho de um tweet

Um outro fator importante na dinâmica de circulação e divulgação de informação dentro da rede social Twitter, é a possibilidade do uso de hashtags. Deste modo, foi feita uma análise de maneira a ilustrar o impacto que uso de *hashtags* tem na disseminação dos tweets, de acordo com os diferentes conteúdos que um tweet pode ter ao longo do tempo, entre 2019 e 2021.

Pela figura 19, em 2019, o uso de hashtags consegue sempre, excluindo os casos para os tópicos "desporto" e "música" em 2019, apresentar melhores resultados, levando a um maior número de partilhas, propensão esta que continuou nos anos seguintes, e onde ainda foi mais relevante a vantagem no uso de hashtags. Em 2020 e 2021 regista-se, uma ainda maior diferença favorável de valores percentuais para tweets com *hashtags* em relação aos sem *hashtags*. Já para os tweets que não usam hashtags e, para a maioria dos tópicos analisados a percentagem que consegue obter tweets não demonstra variar significamente ao longo dos anos.

Figura 19: Percentagem de tweets com partilhas entre tweets com e sem hashtags agrupados por tópico de 2019 a 2021

4.3.4. Análise da dimensão/influência das contas dos retweeters por tópicos entre tweets partilhados e não partilhados

De maneira a compreender melhor o tipo de pessoas que criam conteúdos na rede social, foi feita uma análise à dimensão das contas, isto é, ao seu número de seguidores e o impacto que essa dimensão tem para um dado tweet, seja este ou não partilhado, tendo em conta também a temática dos mesmos.

De uma forma geral, e tendo em conta a figura 20, é notável os tweets que conseguem ser partilhados dentro da rede social, são provenientes de contas que, em média, têm uma quantidade de seguidores substancialmente maior, em comparação com os tweets não partilhados. Não obstante, também se verifica a tendência de que, independentemente do tópico, os tweets não partilhados têm, em média, origem em contas com um número médio de seguidores sempre menor que 2500 seguidores.

Por outro lado, observa-se que ao longo dos anos se registou uma tendência para um aumento na média dos seguidores, de tweets partilhados, enquanto que, para tweets não partilhados, se nota um decréscimo na média de seguidores, desde os seus máximos em 2019. Porém os valores, para tweets partilhados, apresentam, de uma forma geral, uma propensão para subir, sendo que, para alguns casos, verificou-se até, uma recuperação total em 2021, para os valores máximos de 2019.

Já nos extremos dos dados, o tópico "politica" demonstra ser o que é mais mencionado em média por contas com um maior volume de seguidores, tanto para tweets que conseguem ser partilhados como para os que não atingem essa meta.

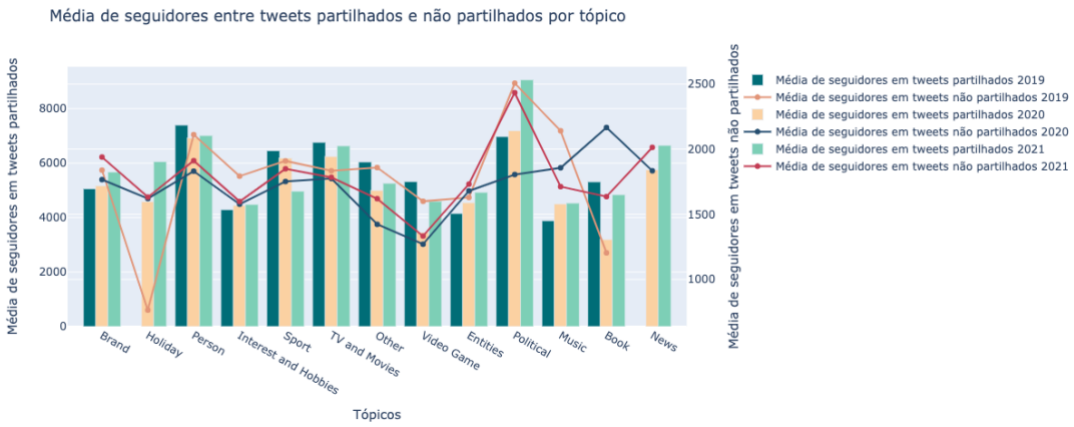

Figura 20: Média de seguidores entre tweets partilhados e não partilhados agrupados por tópicos entre 2019 e 2021

4.4. Análise dos retweeters

4.4.1. Análise dos seguidores dos retweeters por tópico

De maneira a compreender melhor o tipo de pessoas que partilham/propagam conteúdos na rede social (retweeters) foram feitas análises não só à sua presença e participação no Twitter, mas também à maneira como interagem com as diferentes publicações que decidem partilhar com os seus seguidores.

Ao analisar a dimensão das contas que *retweetaram* os tweets recolhidos, na figura 21, verificamos que em média, são contas com uma dimensão considerável, com muitos seguidores e são capazes de alcançar uma grande audiência. Para cada tópico, verifica-se que o número de seguidores, que partilham certos temas é muito diferente. Os tópicos como "filmes e séries" e "pessoas" são os que demonstram ser partilhados por utilizadores com a maior média de seguidores, enquanto que, tópicos como "jogos de digitais" se encontram no lado oposto do espetro, com a menor média de seguidores.

Apesar nas diferenças de média de seguidores entre os diferentes tópicos, nota-se também uma consistência com o passar dos anos. Ainda que para certos tópicos o valor médio de seguidores tenha subido ou descido, as diferenças nunca são muito significativas, excetuando para o tópico "livros" que apresenta um grande pico na média de seguidores que partilham tweets sobre o tema em 2021 (contudo, a súbita subida pode talvez ser explicada, devido há pequena amostra para este tópico (327 tweets em 2021), onde pode existir viés).

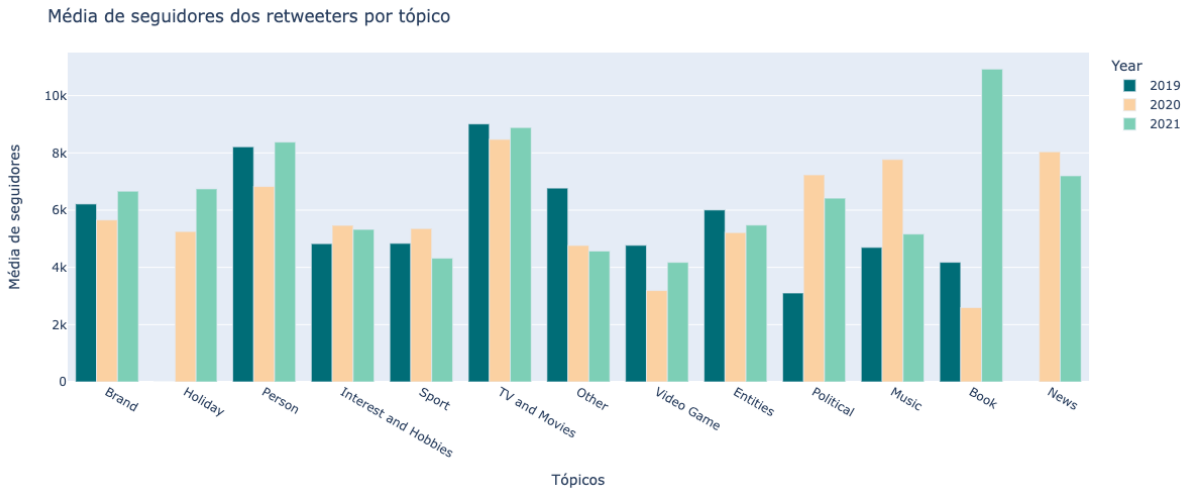

Figura 21: Média de seguidores dos retweeters por tópico

4.4.2. Análise do tempo médio, em dias, para obter todos os retweets por tópico

De forma a melhor compreender os picos de atividade e interações que um dado tweet consegue obter durante o seu ciclo de vida, isto é, desde o momento em que é criado (publicado) até ao momento em que deixa de ser relevante, foi analisado o tempo, em dias, que, em média, é necessário para um tweet conseguir atingir o seu pico de popularidade, por outras palavras, o número máximo de partilhas (retweets) que consegue.

Os retweets conseguidos são em média todos obtidos, em 25 dias desde a publicação de um dado *tweet*. Todavia, certos tópicos apresentam um ciclo de vida mais curto tendo o seu pico de partilhas, em média, menos de 10 dias. Este é o caso dos tópicos "desporto" e "filmes e séries", ou seja, tópicos que demonstram que existe uma grande atividade em volta dos mesmos, o que leva a uma rápida obtenção de partilhas, mas por outro lado origina estagnação rápida na sua disseminação. Em contrapartida, para tópicos como "marcas", o processo de angariação de partilhas é, em média, mais lento que os restantes pares, o que também pode ser um indício que é um tópico que se mantém relevante durante mais tempo.

Não obstante, é notável a tendência para cada vez mais curtos ciclos de vida, dos mais variados tópicos, onde desde 2019 até 2021, a tendência do número médio de dias para obter todas as partilhas tem sido sempre a descer, apresentado quedas significativas de 2019 para 2020, o que pode indicar que os assuntos do "momento" têm cada vez menos tempo na "ribalta".

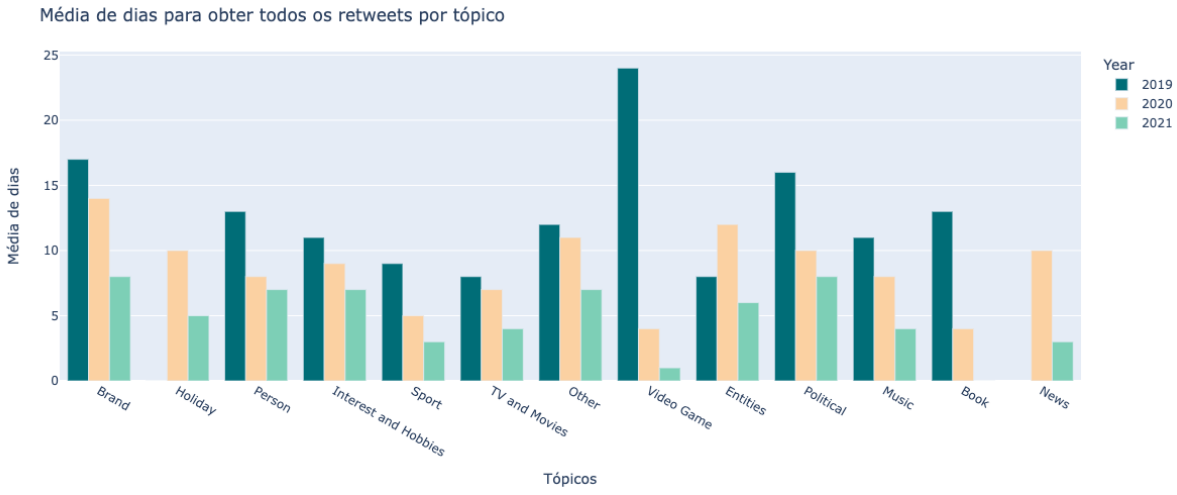

Figura 22: Média de horas para obter as partilhas agrupados por tópico

4.4.3. Análise da antiguidade média dos retweeters por tópico

Sob outra perspetiva, e procurando analisar há quanto tempo as contas que normalmente partilham tweets usam a rede social Twitter, foi feita uma análise à antiguidade destas mesmas contas, abstraindo a análise para o número de anos de uso da plataforma.

Focando na antiguidade das contas de quem partilha os tweets publicados nos mais variados temas, a figura 23 ilustra que estas contas são correspondentes a utilizadores que já frequentam a plataforma há vários anos sendo este fenómeno não só bastante homogêneo entre os diferentes tópicos, como também evidencia uma firme consistência ao longo dos anos.

Assim, revela-se que as contas que partilham tweets, ainda que independentemente do tema em questão, usam por norma a rede social, em média, há aproximadamente 6 anos, sendo apenas registado um valor mais baixo para o tópico “jogos digitais”, talvez pela sua dinamização social e popularidade ser entre as camadas mais jovens da sociedade.

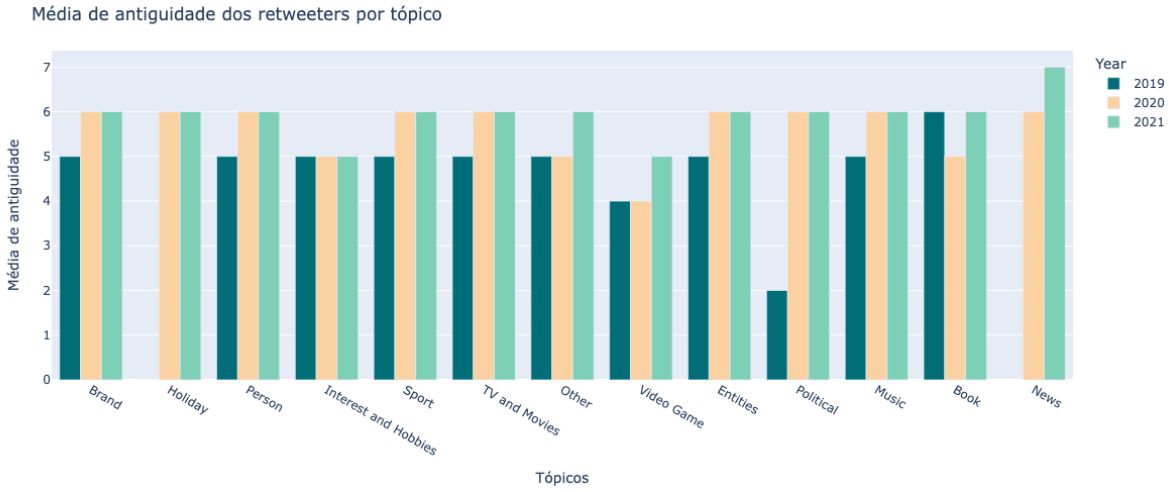

Figura 23: Média da antiguidade das contas dos retweeters agrupados por tópico

4.5. Análise de tweets muito populares (>10 retweets)

4.5.1. Análise do tempo médio, em dias, para obter a primeira metade dos retweets e a sua totalidade

Uma vez que aproximadamente 80% dos tweets recolhidos não conseguiram sequer obter 1 retweet, dos que conseguiram, foi feita uma análise mais dedicada aos que não só foram partilhados pelo menos uma vez, mas que foram partilhados múltiplas vezes, ou seja, tweets "muito populares". Para fazer esta divisão, foi definido o número arbitrário de mínimo 10 retweets, que corresponde a aproximadamente 4,66% dos todos os tweets com pelo menos uma partilha, ou seja, de certa forma uma análise aos outliers que se retiveram no momento da sua remoção, mencionada no capítulo anterior (secção 3.3.6).

Para completar a análise dos utilizadores que partilham tweets e de modo a acrescentar detalhe à análise previamente feita na secção 4.4.1, foi feita uma análise sobre a rapidez com que estes tweets conseguiam obter a primeira metade das partilhas e em quanto tempo restante conseguiam obter os outros 50%. Isto, tendo sempre em conta o tópico em que os *tweets* se inseriam.

A figura 24, ilustra que, independentemente do tópico, a primeira metade das partilhas conseguidas são, de forma geral, obtidas logo durante o primeiro dia da publicação do *tweet*, nunca excedendo, em média, os dois dias, caso se ignore os oito picos registados em certos anos, para certos tópicos. Já a última metade das partilhas, pode levar, em média, até 80 dias a serem obtidas, estando a maioria entre os 10 e os 50 dias, altura em que o tweet atinge o pico da sua disseminação na rede social. Assim, verifica-se que tópicos como "filmes e séries" e "jogos digitais" conseguem muito rapidamente ser partilhados, mas também têm um ciclo de vida de menor que em média não ultrapassa os 20 dias.

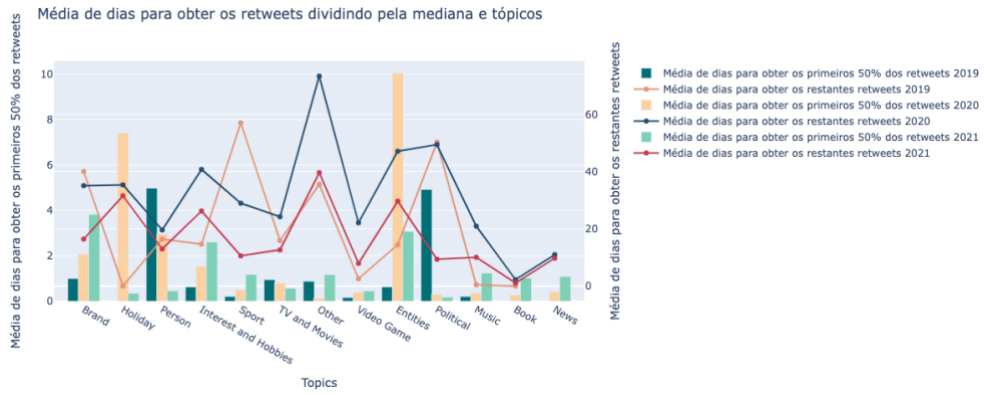

Figura 24: Média do número de horas para obter partilhas separados entre os primeiros e os últimos 50%

4.5.2. Comparação da quantidade média de seguidores dos retweeters entre a primeira e segunda metade

Continuando com a mesma lógica de análise, mas desta vez, focando na dimensão das contas de quem faz a primeira e última metade das partilhas dos tweets "muito populares", a figura 25 demonstra que a primeira metade das partilhas é feita por utilizadores com, em média, uma dimensão ligeiramente superior às contas que fazem a última metade das partilhas. Confirma-se que não existe um padrão ao longo dos anos, pois os valores registados apresentam discrepâncias de ano para ano, mesmo de tópico para tópico.

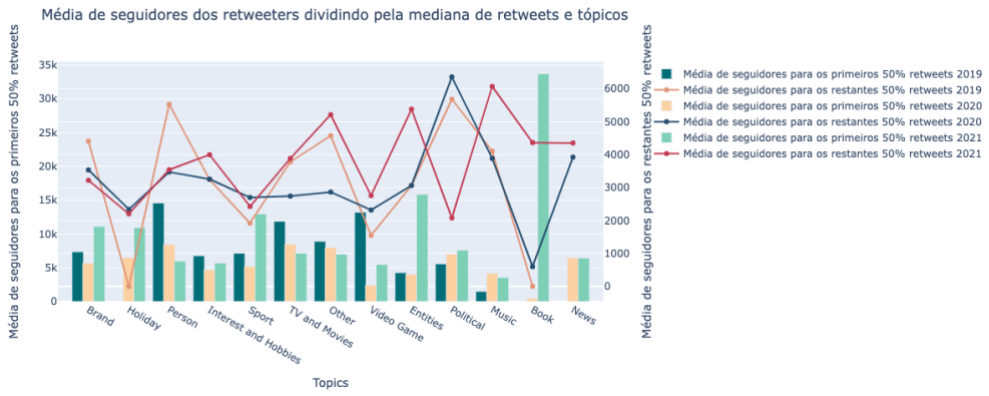

Figura 25: Média dos seguidores do retweeters para obter partilhas separados entre os primeiros e os últimos 50%

4.6. Previsão de tweets populares

4.6.1. Definição do problema e preparação do conjunto de dados

De modo a complementar as análises anteriormente feitas, e com vista a contextualizar de forma prática a informação e características sobre os tweets, foi feita uma modelação recorrendo a algoritmos de aprendizagem supervisionada. Os algoritmos usados são de classificação, pois não se pretende prever o número de partilhas de um tweet, mas sim, se este será ou não partilhado. Desta forma, estes modelos foram treinados recorrendo à classificação previamente feita (se é, ou não, popular) a todos os tweets recolhidos, no capítulo 3 (secção 3.4). Assim, o problema a resolver é um caso de classificação binária.

Neste contexto, os dados recolhidos de 2019 e 2020 foram usados como fonte de treino do modelo selecionado, e os dados de 2021 como material de teste para avaliação do desempenho do modelo na previsão do comportamento da partilha dos tweets. Uma vez que o foco da investigação é compreender também a importância que o tema de cada tweet tem para a sua popularidade dentro da rede social, apenas tweets com tópicos, são retidos e os restantes removidos. Em seguida, e de forma a facilitar o uso de modelos de aprendizagem automática, os conjuntos de dados, o de treino e o de teste, são novamente decompostos, resultando em quatro conjuntos de dados finais. Para esta decomposição e como dados nos conjuntos de validação, ficam os valores previamente definidos pela classificação de "popularidade" mencionada no capítulo 3 (secção 3.4). Já para os dados que o modelo irá usar para treinar, foram selecionadas as seguintes variáveis: as variáveis numéricas *followers*, *following*, *tweet_count* e *seniority*, em conjunto com as variáveis categóricas *topics*, *sentiment*, *hashtags*, *verified*, *day_phase*, *day_of_week* e *month*.

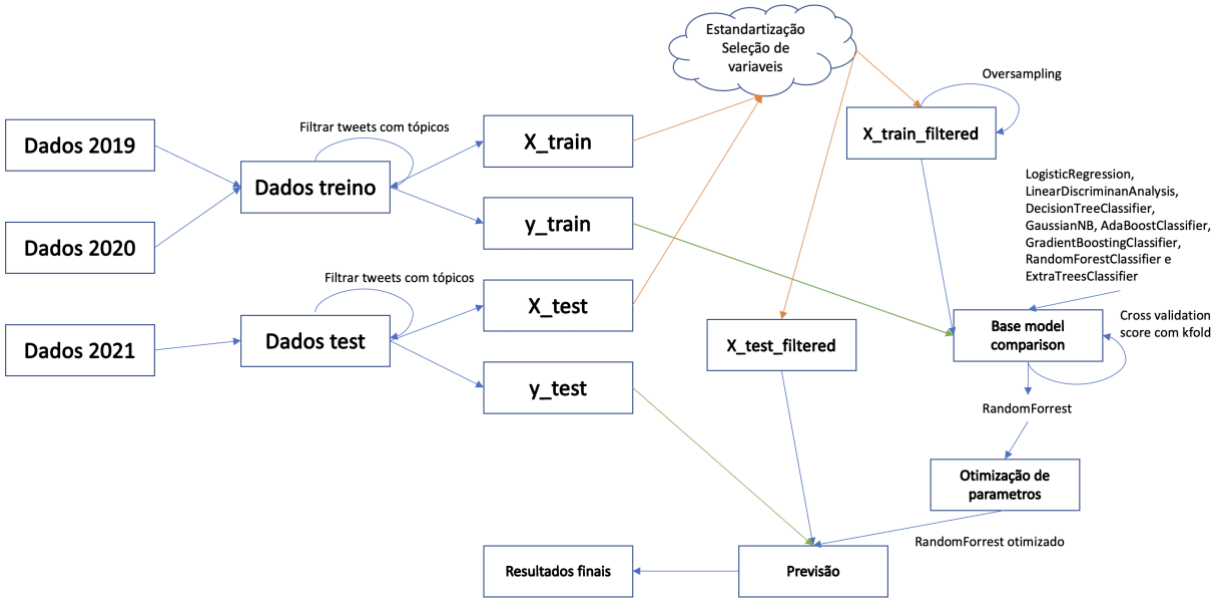

Figura 26: Diagrama do fluxo de desenvolvimento de um modelo de aprendizagem automática

4.6.2. Normalização, seleção de variáveis e sobre amostragem (oversampling)

Devido a certos algoritmos de aprendizagem automática serem mais sensíveis a utilização de variáveis escaladas, em particular para certos modelos, como o *LogisticRegression*, que recorrem ao método de *gradient descent*, e dado que para estes as diferenças entre os intervalos (ranges) das variáveis em estudo, podem causar descidas de diferentes proporções nas convergências de alguns modelos, a utilização das variáveis na mesma escala reduz o impacto. Outros modelos baseados em sistemas de árvores são insensíveis a este tipo de alterações (Bhandari, 2020), mas, apesar disso, e como vários modelos, de ambas as categorias, foram usados numa pré-avaliação, foi aplicado o método de normalização *StandardScaler*, da biblioteca *sklearn*, às variáveis numéricas do conjunto de dados. Para as variáveis categoriais não foi necessária fazer uma normalização dado que estas foram previamente escalas durante o tratamento dos dados, mencionado no capítulo 3 (secção 3.4).

Com as variáveis na mesma escala, e de modo a fazer um estudo de quais as variáveis mais relevantes para a tarefa de previsão a desempenhar, foi feita uma seleção de variáveis separada por dois grupos, seleção de variáveis categoriais e outro de variáveis numéricas, pois devido à natureza das mesmas, diferentes metodologias devem ser aplicadas a cada grupo.

No grupo das variáveis numéricas foram aplicadas quatro análises diferentes, e uma vez que o problema de previsão em estudo é de classificação, a análise das variáveis foi dividida em três partes. Deste modo, e para a procura das variáveis que apresentam melhores resultados, foram aplicadas técnicas como o *recursive feature elimination* (RFE), por filtragem e ainda o recurso a árvores de decisão. Em relação ao método de filtragem, em que é estudada a relação entre as variáveis com a variável a prever (RFE), nesta vertente, métodos estatísticos e de importância de características podem ser usados. Bons candidatos são algoritmos que fazem a sua própria seleção das variáveis, durante o seu treino (Brownlee, 2020). Assim, cada abordagem tem os seus prós e contras, tentando manter uma variedade neste processo para durante a comparação dos resultados de cada uma, a decisão de manter ou remover uma variável ser mais fundamentada.

Posto isto, foi utilizada a *Lasso Regression* que ajuda a reduzir valores nos dados num só ponto central, como uma média, tendo como capacidade a nulificação do impacto de características irrelevantes, e sendo útil para reduzir a possibilidade de overfitting de um modelo (Kumarappan, 2020).

Para complementar a seleção de variáveis com uma análise estatística foi usado o método ANOVA, que estuda a F-Distribuition, ou seja, analisa o impacto da variância, medindo o quão longe um número está da média e de cada número numa variável (Gajawada, 2019).

Como referido anteriormente, foi utilizada a RFE que tem como objetivo a seleção de variáveis onde inicialmente o treino do sistema é feito com todas as variáveis disponíveis. Através de um processo recursivo as variáveis são removidas uma a uma e a importância de cada variável é obtida até chegar ao número mínimo de variáveis desejado. No fim, as variáveis que menos importância demonstraram são removidas (Brownlee, 2020).

Para finalizar a análise das variáveis numéricas, a seleção por sistema de árvores, recorreu-se, mais especificamente, ao algoritmo *extra trees classifier*, que funciona agregando o resultado de múltiplas, não correlacionadas, árvores de decisão combinas numa "floresta", de forma a retornar um resultado a partir dos dados de treino (Semwal et al., 2021). A seleção de variáveis é feita ordenando um conjunto aleatório de características em ordem descendente baseada no indicador de importância *Gini* (mede a probabilidade de uma variável em específico estar incorretamente classificada) para cada variável. Assim, o número das melhores variáveis pode ser escolhido de acordo com a necessidade (Baby et al., 2021).

Depois de feita a seleção das variáveis mais promissoras, apenas as variáveis: *topics_cleaned_entities*, *topics_cleaned_sport*, *topics_cleaned_news*, *sentiment_neutral*, *sentiment_positive*, *hashtags_true*, *verified_true*, *day_phase_dusk*, *day_phase_morning*, *month_december*, *followers*, *following*, *seniority* e *tweet_count* foram selecionadas sendo que algumas delas, apenas representam as partes mais relevantes das variáveis categóricas previamente decompostas recorrendo ao processo de OneHotEncoder. Tendo em ainda em conta a grande desproporcionalidade da variável que se pretende classificar, a popularidade, isto é, se um tweet consegue, ou não, obter retweets, recorreu-se a um método de balanceamento da amostra de dados, em particular o SMOTE. Contudo, o uso de um método de oversampling, como o SMOTE aumenta a probabilidade de overfitting (Chawla et al., 2002) e também o tempo computacional requerido para treinar os modelos, uma vez que o tamanho da amostra aumenta. Tendo em conta a falta de dados correspondentes a casos "populares", em comparação com o seu inverso, o seu uso permitia não ter de descartar dados dos tweets "não populares", para que a amostra de dados ficasse equilibrada. Para diferentes modelos, a técnica de oversampling consegue melhores desempenhos e melhores resultados para diferentes métricas de avaliação (Mohammed et al., 2020). Com esta técnica, a amostra de dados continha inicialmente 322.530 tweets não populares e 60.677 populares e ficou com 322.530 não populares e 322.530 populares.

4.6.3. Escolha do modelo, otimizações e resultados

Após a decisão sobre as variáveis finais a utilizar de acordo com a seleção, anteriormente mencionada, é feita uma comparação entre oito modelos de *machine learning* na sua configuração base, *LogisticRegression*, *LinearDiscriminanAnalysis*, *DecisionTreeClassifier*, *GaussianNB, AdaBoostClassifier*, *GradientBoostingClassifier*, *RandomForestClassifier e ExtraTreesClassifier*. Para efetuar a comparação, é usado o sistema de *cross validation* acompanhado do uso de *kfold*. Esta metodologia procura testar o desempenho de cada modelo com diferentes excertos dos dados de treino, dividindo os dados em "n" número de *folds*. Este sistema ajuda a evitar situações de *overfitting*. Apesar dos benefícios, tem como lado negativo, ser um teste computacionalmente exigente, que se traduz em mais tempo de processamento e com a agravante de aumentar consoante a quantidade de dados a processar.

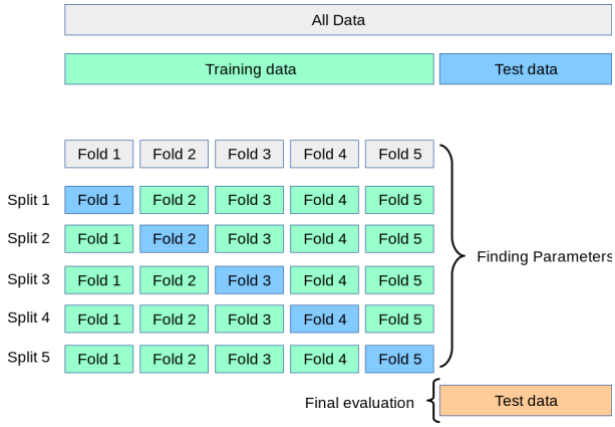

Figura 27: Representação do sistema de cross validation score com cinco folds[3]

Da avaliação e posterior comparação referida, como pode ser observado na figura 28, é apresentada a comparação das diferentes precisões obtidas pelo sistema de *cross validation* com cinco *folds* e escolhido o modelo com melhor média. Os modelos do tipo *ensemble* conseguiram melhores resultados e destes, os modelos que recorrem a métodos de *bagging* como o *RandomForest* e o *ExtraTress* foram os que obtiveram os melhores resultados face aos modelos que usam métodos de *boosting*. Assim o modelo *ExtraTrees* é o que apresenta melhor desempenho médio e, foi o escolhido.

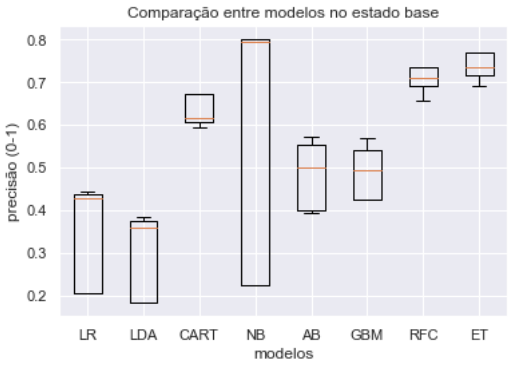

Figura 28: Resultados da precisão dos oito modelos no estado base recorrendo a cross validation

Deste modo, e a fim de garantir o melhor desempenho possível deste modelo para os dados em estudo, é feita uma *GridSearchCV*, que permite uma pesquisa exaustiva para

[3] Imagem retirada do site: https://scikit-learn.org/stable/modules/cross_validation.html

um dado conjunto de parâmetros, conseguindo aferir quais os melhores de uma lista previamente fornecida (figura 29).

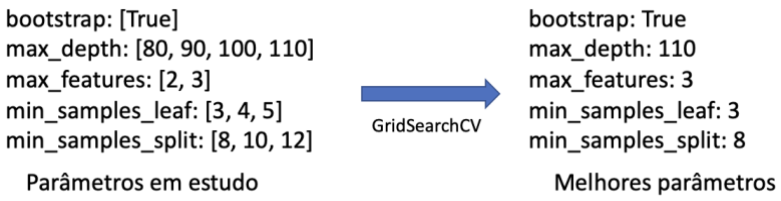

Figura 29: Uso do método de GridSearchCV para encontrar os melhores parâmetros para o modelo

Depois de identificado o melhor modelo e quais os melhores parâmetros para a sua configuração, este é testado com os dados de 2021, guardados como dados de teste, no início do processo. Assim, e verificando os resultados obtidos pelo modelo, na tabela 1, o modelo escolhido conseguiu prever se um *tweet* ia ou não ser "popular", mostrando um nível de exatidão de 76%. Também se verifica que dado a maior abundância de dados sobre tweets não populares, o modelo conseguiu níveis de precisão mais elevados para esta classe, de 89% face a apenas 34% para tweets populares. A discrepância é evidenciada pelos valores de suporte, que ilustra a dimensão de cada amostra.

Tabela 1: Resultados do desempenho do modelo criado testado com dados de 2021

	Precisão	**Cobertura**	**F1-score**	**Suporte**
Não popular	0,89	0,81	0,85	145295
Popular	0,34	0,51	0,41	28813
Exatidão		0,76		

Por tudo isto, os resultados obtidos, demonstram ser satisfatórios e permitem prever, com uma alguma precisão, se um tweet irá ou não ser popular com base nas suas características. Contudo, os resultados também revelam que o modelo tem mais dificuldade em identificar tweets que são populares o que pode ser explicado pela grande discrepância na sua representação no conjunto de dados recolhidos, ainda que se tenha recorrido ao uso de métodos de oversampling de forma a minimizar o impacto. Estes resultados deixam ainda uma margem de melhoria a explorar em trabalhos futuros.

4.7. Caso de estudo: Covid-19

Uma vez que o período analisado neste estudo, de 2019 a 2021, inclui o começo, bem como, a maturação do surgimento do covid-19, foi feito um caso de estudo sobre os conceitos mencionados nas secções anteriores deste mesmo capítulo, à luz do impacto que a disseminação desta epidemia teve também nas redes sociais, analisando como surgiu e se maturou.

De modo a realizar esta análise, foi feita uma pesquisa pela existência de palavras-chave no corpo de texto de cada tweet recolhido durante o ano de 2020, ano em que oficialmente foi declarado o começo da pandemia nos Estados Unidos da América (país de onde os tweets deste estudo foram recolhidos). A figura 30 exibe as palavras-chaves usadas para efetuar a pesquisa e estas foram retiradas de um estudo começado em 2020, que pretende recolher tweets sobre os acontecimentos relacionados com covid-19 no Twitter (E. Chen et al., 2020). De forma a comparar a efetividade do uso de palavras-chave para identificação de tweets sobre covid, o mesmo processo de procura foi também usado nos tweets de 2019 e 2021, contudo, como a pandemia só teve o seu início mais mediático já em 2020, o esperado é um número bastante reduzido de tweets em 2019, com maior incidência no fim desse ano. Assim, com este método, foram identificados 285 tweets sobre covid-19 em 2019, 3983 em 2020 e 1581 em 2021.

```
['Coronavirus', 'Corona', 'CDC', 'Ncov', 'Wuhan', 'Outbreak', 'China', 'Koronavirus',
 'Wuhancoronavirus', 'Wuhanlockdown', 'N95', 'Kungflu', 'Epidemic', 'Sinophobia',
 'Covid-19', 'Corona virus', 'Covid19', 'Sars-cov-2', 'COVID-19', 'COVD', 'Pandemic',
 'Coronapocalypse', 'CancelEverything', 'Coronials', 'SocialDistancing', 'Panic buying',
 'DuringMy14DayQuarantine', 'Panic shopping', 'InMyQuarantineSurvivalKit', 'chinese virus',
 'stayhomechallenge', 'DontBeASpreader', 'lockdown', 'shelteringinplace', 'staysafestayhome',
 'trumppandemic', 'flatten the curve', 'GetMePPE', 'covidiot', 'epitwitter', 'Pandemie',
 'PneumoniaWuhan', 'CoronaVirusInfo', 'V2019N', 'CDCemergency', 'CDCgov', 'WHO', 'HHSGov', 'NIAIDNews']
```

Figura 30: Palavras-chave utilizadas para identificar tweets sobre covid-19

Após o processo de classificação dos tweets, foi feita uma análise da média de partilhas e gostos ao longo do ano de 2020, como ilustra a figura 31, onde é possível verificar que existiram dois grandes picos de atividade, em março e em setembro, que de acordo com a linhagem de eventos sobre a doença em 2020 (AJMC Staff, 2021) coincidem com a declaração do covid-19 como uma pandemia pela WHO a 11 de março de 2020 e com a divulgação da entrada de, à data, várias potenciais vacinas na fase três dos testes clínicos em setembro. Estes eventos podem explicar a acentuada quantidade de partilhas durante

estes meses e até o aumento do número médio de gostos nos tweets sobre as informações de setembro, pois seriam informações muito positivas para o combate à doença.

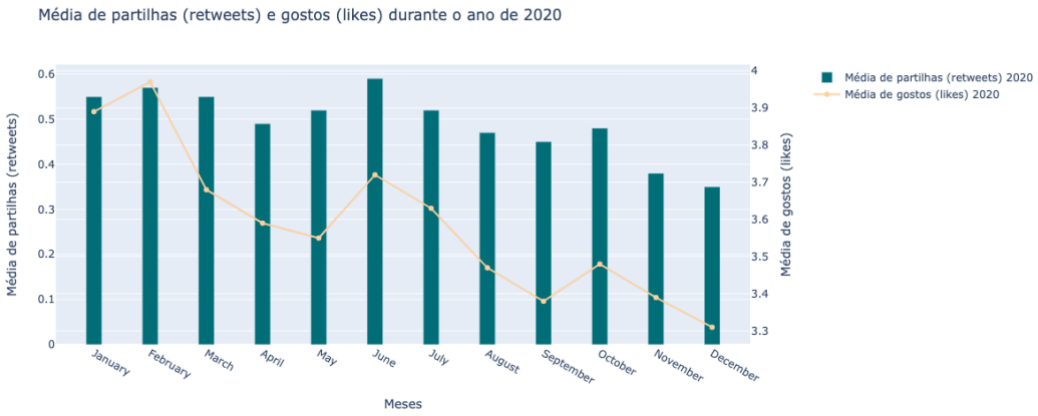

Figura 31: Média de partilhas e gostos durante o ano de 2020

Nota-se também, uma clara diferença no número médio de partilhas e gostos entre tweets sobre covid-19 e os restantes tópicos, diferença essa, que demonstra que tweets sobre covid-19 durante 2020 foram sempre mais populares a partir do mês de janeiro, que é quando a doença começa a ser falada pelas massas em maior quantidade.

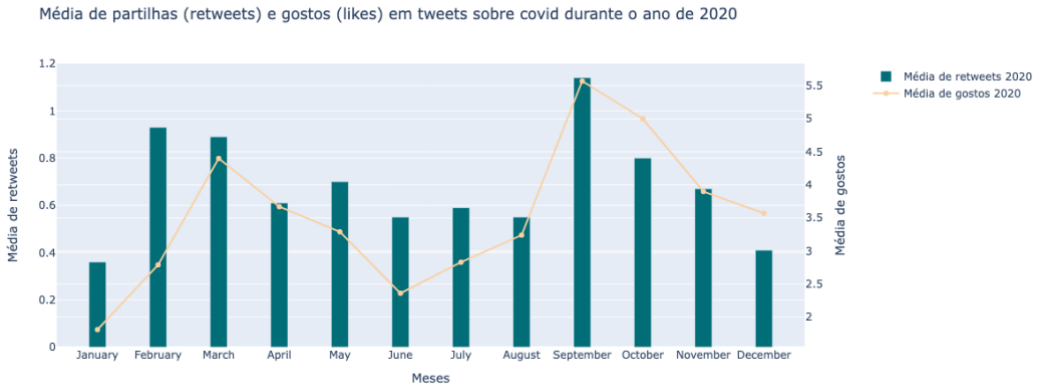

Figura 32: Média de partilhas e gostos para tweets sobre covid-19 em 2020

O facto de, de forma geral, ter sido registada uma menor atividade de interação (partilhas e gostos) com os tweets em 2020, como constatado nas secções anteriores deste capítulo, contraria a popularidade ser superior em tweets sobre covid, o que parece demonstrar que o aparecimento do covid-19 e o seu impacto nas redes socias, não explica a queda de atividade durante este ano.

Capítulo 5 – Conclusões e trabalho futuro

Neste capítulo, discutem-se os resultados da investigação realizada. Para além disso, são abordadas as limitações da investigação e propostas novas ideias para continuação do trabalho desenvolvido.

Neste estudo foi implementado um processo de recolha e tratamento de dados históricos, para o período temporal do Twitter ao longo de três anos, com o intuito da criação de um conjunto de dados representativo de cada ano recolhido, bem como as diferentes fases de cada dia. O estudo tem como objetivo analisar o comportamento que os tweets têm, desde a sua criação até ao pico da sua popularidade, dentro da rede social, para além disso, categorizar quem interage e torna tweets populares e o modo como o faz. Este processo também teve em conta o tópico e sentimento expresso em cada tweet, de modo a estudar como a temática do mesmo influencia a sua popularidade dentro da rede social.

De forma a alcançar este objetivo, foi usada a API do Twitter e concebidas duas maneiras de recolher os dados, uma focada em recolher os tweets representativos do período temporal ao longo do ano, e outra que visa recolher de uma maneira mais dinâmica, todos os utilizadores que partilharam os tweets recolhidos através da metodologia proposta neste estudo. Uma vez que, a API não fornece um modo fácil para a recolha de dados, quando em causa estão grandes volumes de tweets. Foi concluído, que uma boa maneira de alcançar esta recolha é usar o método de recolha de tweets da API, mas com um critério de pesquisa que, combina o url do tweet original com o nome do utilizador que criou o tweet. Este processo, ainda que falhando em recolher os retweeters de todos os tweets, conseguiu coletar com sucesso a maioria de tweets relevantes.

A partir dos dados recolhidos foi possível retirar conclusões sobre os momentos mais oportunos de partilhar conteúdo no Twitter dependendo do tema do tweet. No que diz respeito à fase do dia, verificou-se que a hora de almoço demonstra ser o melhor momento para partilhar conteúdo, seguida da noite. Quanto ao dia da semana, os dias a meio da mesma, aparentam ser a altura mais propícia para a sua disseminação. Para além disso, o sentimento de um tweet quando aliado a um dado tópico, demonstra que, tweets com sentimento positivo conseguem ter uma maior relevância na disseminação do seu conteúdo, quando comparados com tweets sobre o mesmo tema, mas que transmitem

sentimentos negativos ou neutros. Dos tópicos analisados, e mesmo ao longo dos anos, a maioria dos tópicos apresentava sempre melhores estatísticas de disseminação quando o sentimento era positivo. Para além disso, o uso de "hashtags" também demonstra ter benefício, em comparação com a sua omissão, em que para todos os tópicos, o nível de popularidade conseguiu ser sempre superior, quando os tweets recorriam à sua utilização.

Quando analisados os utilizadores por de trás dos tweets, e comparando tweets que foram partilhados, com os que não conseguiram partilhas, os que foram partilhados provêm de utilizadores com um grande número médio de seguidores face aos tweets sem partilhas, demonstrando uma diferença considerável quando comparando as duas vertentes. Também se verificou que, os tweets partilhados, foram disseminados por utilizadores que, em média, já frequentavam o Twitter há cinco e seis anos. Estes retweeters são utilizadores com perfis diferentes dependendo do tópico em análise, ou seja, para tópicos como "TV e filmes" os retweeters deste tema, em média, apresentam ter contas com aproximadamente dez mil seguidores, mas por outro lado, para tópicos como "jogos digitais" o número de seguidores ronda os quatro mil. Os retweeters de uma forma geral, demonstram ser utilizadores com um grande número de seguidores, em média, na casa dos milhares. De outra forma, reconhece-se uma tendência decrescente no número médio de dias que um tweet necessita, para obter todos os retweets, tanto que em 2021, já menos de nove dias são necessários. Assim, também se verifica que a "esperança média de popularidade" de um tweet tem vindo a diminuir.

Para complementar a análise mencionada, foi também feito um estudo aos tweets muito populares (com pelo menos dez retweets) de forma a compreender melhor se existem diferenças que não só levam um tweet a conseguir, pelo menos uma partilha (que dada a pequena quantidade recolhida, demonstra ser algo por si só difícil), mas também, se existem características que impulsionam os tweets a irem além, e conseguirem reunir múltiplas partilhas. Dentro desta vertente, conclui-se que os tweets muito populares, conseguem rapidamente obter a primeira metade dos seus retweets, sendo que esta primeira metade é concretizada por contas que, em média, têm um maior número de seguidores do que a restante metade.

Foi também desenvolvido um modelo que tenta prever se um dado tweet vai ou não obter partilhas. Para a criação deste modelo, os dados recolhidos de 2019 e 2020 alimentaram o modelo para a sua fase de treino e os dados de 2021 foram usados como base de teste para medir a precisão do modelo gerado. De modo a conceber o modelo, e

para além do tratamento feito aos dados recolhidos mencionado no capítulo 3 (secção 3.4), foi feita uma normalização dos dados, seguida de uma seleção das melhores variáveis e finalizando com a aplicação de um método de *oversampling* para a expansão dos poucos dados de tweets com partilhas previamente recolhidos. Em seguida foi feito um estudo de qual o modelo que melhor se adaptava ao problema, concluindo com a identificação dos melhores híper-parâmetros para o modelo selecionado, de forma a garantir que este se encontrava o mais otimizado possível.

Os resultados que o modelo gerado alcançou foram satisfatórios, mas demonstram que ainda existe espaço para melhorias na previsão, bem como acentuam a dificuldade que é modelar o comportamento de partilha dentro das redes socias.

O sistema desenvolvido em *python*, pode ser melhorado, otimizando o método de recolha de dados, para que a amostra recolhida seja o mais representativa possível. Para além disso, a parte automática de construção de uma representação visual dos dados recolhidos, pode ser tornada numa dashboard interativa hospedada num *website*, facilitando o acesso à informação e o estado atual da disseminação da informação. Outra possibilidade de melhoria, seria a implementação de melhores heurísticas de categorização do tema e tópico do tweet, bem como usando mais dados de treino para preparar o modelo de previsão.

Por fim, um estudo mais aprofundado sobre o comportamento dos retweeters de um tweet pode também trazer melhores resultados, aquando da previsão dos mesmos, passando por melhorar a metodologia de recolha destes dados, pois a apresentada neste estudo ainda deixa uma margem para otimização. Para além disso, e uma vez que os dados recolhidos são provindos de utilizadores dos E.U.A., outros *corpus* de dados podem dar resultados diferentes, devido ao comportamento dos utilizadores noutros países poder ser diferente.

Referências

AJMC Staff. (2021, January 1). *A Timeline of COVID-19 Developments in 2020*. https://www.ajmc.com/view/a-timeline-of-covid19-developments-in-2020

Antonakaki, D., Fragopoulou, P., & Ioannidis, S. (2021). A survey of Twitter research: Data model, graph structure, sentiment analysis and attacks. *Expert Systems with Applications*, *164*. https://doi.org/10.1016/j.eswa.2020.114006

Baby, D., Devaraj, S. J., Hemanth, J., & Anishin Raj, M. M. (2021). Leukocyte classification based on feature selection using extra trees classifier: A transfer learning approach. *Turkish Journal of Electrical Engineering and Computer Sciences*, *29*, 2742–2757. https://doi.org/10.3906/elk-2104-183

Benevenuto, F., Magno, G., Rodrigues, T., & Almeida, V. (2010). *Detecting Spammers on Twitter*.

Bhandari, A. (2020). *Feature Scaling | Standardization Vs Normalization*. https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/

Binns, A. (2012). Don't feed the trolls!: Managing troublemakers in magazines' online communities. *Journalism Practice*, *6*(4), 547–562. https://doi.org/10.1080/17512786.2011.648988

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., van Bavel, J. J., & Fiske, S. T. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(28), 7313–7318. https://doi.org/10.1073/pnas.1618923114

Bright, J. (2018). Explaining the emergence of political fragmentation on social media: The role of ideology and extremism. *Journal of Computer-Mediated Communication*, *23*(1), 17–33. https://doi.org/10.1093/jcmc/zmx002

Brownlee, J. (2020a, August 20). *How to Choose a Feature Selection Method For Machine Learning*. https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/

Brownlee, J. (2020b, August 28). *Recursive Feature Elimination (RFE) for Feature Selection in Python*. https://machinelearningmastery.com/rfe-feature-selection-in-python/

Can, U., & Alatas, B. (2017). Big social network data and sustainable economic development. *Sustainability (Switzerland)*, *9*(11). https://doi.org/10.3390/su9112027

Chawla, N. v, Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. In *Journal of Artificial Intelligence Research* (Vol. 16).

Chen, E., Lerman, K., & Ferrara, E. (2020). Tracking social media discourse about the COVID-19 pandemic: Development of a public coronavirus Twitter data set. *JMIR Public Health and Surveillance*, *6*(2). https://doi.org/10.2196/19273

Chen, X., & Li, J. (2019). Community detection in complex networks using edge-deleting with restrictions. *Physica A: Statistical Mechanics and Its Applications*, *519*, 181–194. https://doi.org/10.1016/j.physa.2018.12.023

Chu, Z., Widjaja, I., & Wang, H. (2012). *LNCS 7341 - Detecting Social Spam Campaigns on Twitter*.

Gajawada, S. (2019, October 19). *ANOVA for Feature Selection in Machine Learning | by sampath kumar gajawada | Towards Data Science*.

https://towardsdatascience.com/anova-for-feature-selection-in-machine-learning-d9305e228476
Halpern, D., Valenzuela, S., & Katz, J. E. (2017). We Face, I Tweet: How Different Social Media Influence Political Participation through Collective and Internal Efficacy. *Journal of Computer-Mediated Communication*, *22*(6), 320–336. https://doi.org/10.1111/jcc4.12198
Hodas, N. O., & Lerman, K. (2013). *The Simple Rules of Social Contagion*. https://doi.org/10.1038/srep04343
Hubert, R. B., Estevez, E., Maguitman, A., & Janowski, T. (2020). Analyzing and Visualizing Government-Citizen Interactions on Twitter to Support Public Policy-making. *Digital Government: Research and Practice*, *1*(2), 1–20. https://doi.org/10.1145/3360001
Hung, M., Lauren, E., Hon, E. S., Birmingham, W. C., Xu, J., Su, S., Hon, S. D., Park, J., Dang, P., & Lipsky, M. S. (2020). Social network analysis of COVID-19 sentiments: Application of artificial intelligence. *Journal of Medical Internet Research*, *22*(8). https://doi.org/10.2196/22590
Knobloch-Westerwick, S., & Meng, J. (2009). Looking the other way: Selective exposure to attitude-consistent and counterattitudinal political information. *Communication Research*, *36*(3), 426–448. https://doi.org/10.1177/0093650209333030
Kolomeets, M., Benachour, A., Baz, E., Chechulin, A., Strecker, M., Kotenko, I., & el Baz, D. (2019). *Ref-erence architecture for social networks graph analysis tool*. *10*(4), 109–125. https://doi.org/10.22667/JOWUA.2019.12.31.109ï
Kumarappan, S. (2020, August 16). *Feature Selection by Lasso and Ridge Regression-Python Code Examples | by Sabarirajan Kumarappan | Medium*. https://medium.com/@sabarirajan.kumarappan/feature-selection-by-lasso-and-ridge-regression-python-code-examples-1e8ab451b94b
Li, X., Zhou, S., Liu, J., Lian, G., Chen, G., & Lin, C. W. (2019). Communities detection in social network based on local edge centrality. *Physica A: Statistical Mechanics and Its Applications*, *531*. https://doi.org/10.1016/j.physa.2019.121552
Liu, H., Christiansen, T., Baumgartner, W. A., & Verspoor, K. (2012). BioLemmatizer: A lemmatization tool for morphological processing of biomedical text. *Journal of Biomedical Semantics*, *3*(1). https://doi.org/10.1186/2041-1480-3-3
Mohammed, R., Rawashdeh, J., & Abdullah, M. (2020). Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. *2020 11th International Conference on Information and Communication Systems, ICICS 2020*, 243–248. https://doi.org/10.1109/ICICS49469.2020.239556
Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302
scikit-learn. (n.d.). Retrieved August 22, 2022, from https://scikit-learn.org/stable/
Semwal, V. B., Lalwani, P., Mishra, M. K., Bijalwan, V., & Chadha, J. S. (2021). An optimized feature selection using bio-geography optimization technique for human walking activities recognition. *Computing*, *103*(12), 2893–2914. https://doi.org/10.1007/s00607-021-01008-7
Shin, J., Jian, L., Driscoll, K., & Bar, F. (2018). The diffusion of misinformation on social media: Temporal pattern, message, and source. *Computers in Human Behavior*, *83*, 278–287. https://doi.org/10.1016/j.chb.2018.02.008

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake News Detection on Social Media. *ACM SIGKDD Explorations Newsletter*, *19*(1), 22–36. https://doi.org/10.1145/3137597.3137600

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Waszak, P. M., Kasprzycka-Waszak, W., & Kubanek, A. (2018). The spread of medical fake news in social media – The pilot quantitative study. *Health Policy and Technology*, *7*(2), 115–118. https://doi.org/10.1016/j.hlpt.2018.03.002

Yang, A., Choi, I. M., Abeliuk, A., & Saffer, A. (2021). The Influence of Interdependence in Networked Publics Spheres: How Community-Level Interactions Affect the Evolution of Topics in Online Discourse. *Journal of Computer-Mediated Communication*. https://doi.org/10.1093/jcmc/zmab002

Yang, K. C., Varol, O., Davis, C. A., Ferrara, E., Flammini, A., & Menczer, F. (2019). Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies*, *1*(1), 48–61. https://doi.org/10.1002/hbe2.115